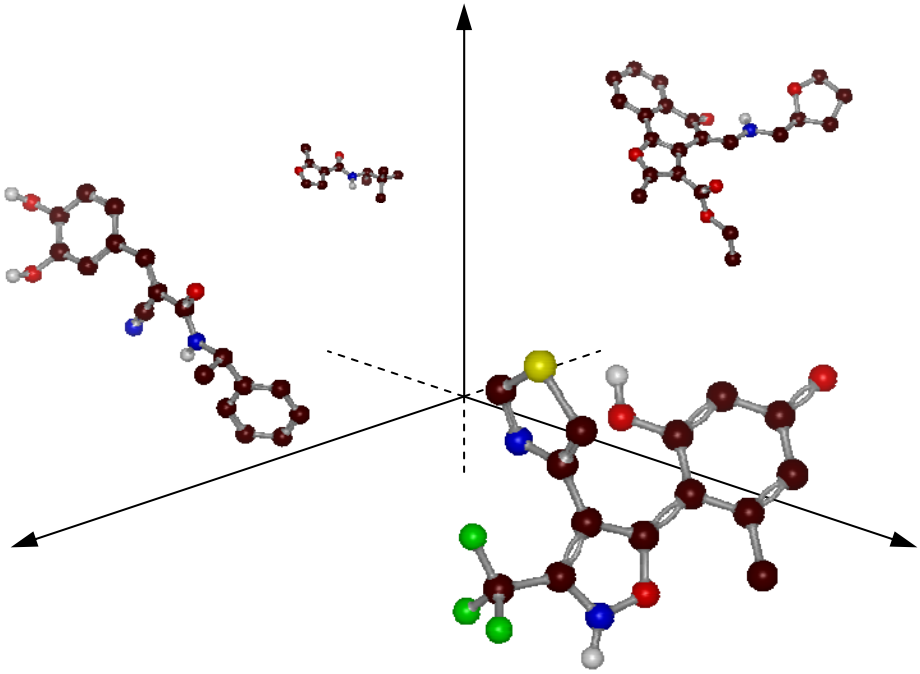
Spatial Statistics Methods for the Analysis of Chemical Datasets in Virtual Screening Validation Experiments



Von der Fakultät für Lebenswissenschaften

der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

genehmigte

Dissertation

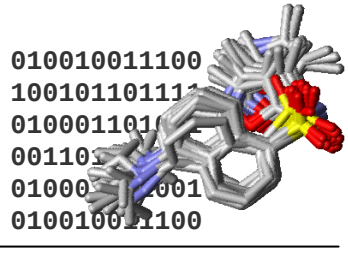
von

Sebastian Georgios Rohrer

aus

Nürnberg

Braunschweig 2008



Spatial Statistics Methods for the Analysis of Chemical Datasets in Virtual Screening Validation Experiments

Von der Fakultät für Lebenswissenschaften

der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

genehmigte

DISSERTATION

von: Sebastian Georgios Rohrer

aus: Nürnberg

1. Referent:	Prof. Dr. Knut Baumann
2. Referent:	Prof. Dr. Hermann Wätzig
eingereicht am:	10.11.2008
mündliche Prüfung (Disputation) am:	17.12.2008
Druckjahr:	2008

Vorveröffentlichungen der Dissertation

Teilergebnisse dieser Arbeit wurden mit Genehmigung der Fakultät für Lebenswissenschaften, vertreten durch den Mentor dieser Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Originalpublikationen in Fachjournalen mit Peer-Review

Rohrer, S.G.; Baumann, K.

Impact of Benchmark Data Set Topology on the Validation of Virtual Screening Methods: Exploration and Quantification by Spatial Statistics.

J. Chem. Inf. Model., 2008, 48, 704-71

Rohrer, S.G.; Baumann, K.

Maximum Unbiased Validation (MUV) Datasets for Virtual Screening Based on PubChem Bioactivity Data

J. Chem. Inf. Model., im Druck

Vorträge auf Fachkongressen

Rohrer, S.G., Baumann K.

Maximum Unbiased Validation (MUV) Datasets for Benchmarking of Ligand-Based Virtual Screening Techniques

Jahrestagung der Deutschen Pharmazeutischen Gesellschaft , Bonn, 2008

Rohrer, S.G.

Exploring Benchmark Dataset Bias in Ligand Based Virtual Screening

Informa Life Sciences: Design and Synthesis of Quality Compound Libraries, München, 2007

Rohrer, S.G.

Fair Play in Virtual Screening: Ruling Out Database Composition as a Critical Factor in the Validation of Ligand based Virtual Screening Methods

20. Darmstädter Molecular Modelling Workshop, Erlangen, 2006

Posterbeiträge

Rohrer S.G., Baumann K.

Fair Pay in Virtual Screening: Ruling Out the Influence of Database Composition on the Validation of Virtual Screening Techniques.

Molecular Modelling 2006 (General Meeting of the MGMS Asia-Pacific Chapter), Perth, Australia, 2006

Rohrer S.G., Baumann K.

Fair Pay in Virtual Screening: Revealing the Influence of Database Composition on the Validation of Virtual Screening Techniques.

Summer School on Biomolecular Simulations, European Molecular Biology Organization (EMBO), Institute Pasteur, Paris, France, 2006

Rohrer S.G., Baumann K.

The Real Revenue from Virtual Screening: The Impact of Benchmark Dataset Bias on Figures of Merit.

2nd German Conference on Chemoinformatics: 20. CIC-Workshop, Goslar, 2006

Rohrer S.G., Baumann K.

Exploring benchmark dataset bias in ligand based virtual screening using Self-Organizing Maps.

National Meeting of the American Chemical Society, Chicago, USA, 2007

Rohrer S.G., Baumann K.

Exploring Benchmark Dataset Bias in Ligand Based Virtual Screening using Self-Organizing Maps.

4th Joint Sheffield Conference on Chemoinformatics, Sheffield, UK, 2007

Rohrer S.G., Baumann K.

Exploring Benchmark Dataset Bias in Ligand based Virtual Screening.

3rd German Conference on Chemoinformatics: 21. CIC-Workshop, Goslar, 2007

Rohrer, S.G.

Exploring Benchmark Dataset Bias in Ligand Based Virtual Screening

Informa Life Sciences: Design and Synthesis of Quality Compound Libraries, München, 2007

Rohrer S.G., Baumann K.

Maximum Unbiased Validation (MUV) Datasets for Virtual Screening by PubChem Based Chemogenomics Data Mining.

8th International Conference on Chemical Structures, Noordwijkerhout, The Netherlands, 2008

Rohrer S.G., Baumann K.

MUVing SAFEly – Scale for Assessing Figures of Effectiveness (SAFE) in Virtual Screening using Maximum Unbiased Validation (MUV) Datasets.

4th German Conference on Chemoinformatics: 22. CIC-Workshop, Goslar, 2008

Acknowledgements - Danksagungen

Diese Arbeit entstand am Institut für Pharmazie und Lebensmittelchemie der Bayerischen Julius-Maximilians Universität Würzburg sowie am Insitut für Pharmazeutische Chemie der Technischen Universität Carolo-Wilhelmina zu Braunschweig unter der Anleitung von

Professor Dr. Knut Baumann.

Lieber Knut, ich danke dir für $3^1/_2$ spannende, interessante und lehrreiche Jahre, die mich oft an meine Grenzen geführt haben. Du hast es nie versäumt, mir Wege aufzuzeigen, die mich immer wieder über diese Grenzen hinaus zu neuen Horizonten geführt haben.
Das Schönste, was wir erleben können, ist das Geheimnisvolle. Es ist das Grundgefühl, das an der Wiege von wahrer Kunst und Wissenschaft steht. Wer es nicht kennt und sich nicht wundern, nicht mehr staunen kann, der ist sozusagen tot und sein Auge erloschen. - Albert Einstein

Dank gilt auch Herrn Prof. Dr. Hermann Wätzig für die Übernahme des Koreferats, sowie Frau Prof. Dr. Heike Faßbender und Herrn Prof. Dr. Conrad Kunick, die als Prüfer in der Disputation fungierten.

Ich möchte außerdem allen Mitgliedern der Arbeitskreise Kunick, Wätzig und Baumann für die schöne Zeit in Braunschweig danken. Ganz besonders möchte ich mich hier bei Anja Becker, Markus Kossner, Ulrike Schmid, Christina Anthes, Jan Dreher und Steffi Lütge für die Kaffeepausen, Parties, Konferenzbesuche, aber auch für die inspirierenden wissenschaftlichen Diskussionen danken, ohne die diese Arbeit nicht möglich gewesen wäre.

"If I have seen a little further it is by standing on the shoulders of Giants."
Isaac Newton

Gewidmet meiner Familie.
Meinen Wurzeln. Meiner Zukunft.
Sebastian Rohrer, Braunschweig im November 2008

Contents

Vorveröffentlichungen der Dissertation iv

Acknowledgements - Danksagungen vii

Abstract xiii

Zusammenfassung xiv

1 Introduction 1

1.1 Virtual Screening in the Process of Drug Discovery 1

1.1.1 Screening: An Entry-Point to Drug Discovery 1

1.1.2 Experimental High-Throughput Screening (HTS) 2

1.1.2.1 Common HTS Assay Formats 4

1.1.2.2 PubChem: HTS Data for the Public Sector 5

1.1.3 Virtual Screening (VS) 6

1.1.3.1 Structure Based Virtual Screening (SBVS) 6

1.1.3.2 Ligand Based Virtual Screening (LBVS) 8

1.2 Validation of Virtual Screening Techniques 14

1.2.1 Objectives of Validation Experiments 14

1.2.2 Validation Procedures and Figures of Merit (FoM) 15

1.3 Benchmark Datasets 19

1.3.1 Available Benchmark Datasets for VS Validation 19

1.3.2 Impact of Dataset Composition on Validation Results 20

1.3.3 Chemical Space 21

1.3.4 Benchmark dataset topology 23

2 Impact of Dataset Topology on VS Validation 24

2.1 Objectives 24

2.2 Methods 25

2.2.1 Methodological Strategy 25

2.2.2 Datasets 25

2.2.3 Descriptors 27

2.2.4 Sub-Samples with Defined Topology 29

2.2.5 Retrospective Virtual Screening Simulations 30

2.2.6 Figures of Merit (FoM) for Virtual Screening Performance 30

2.2.7 Variance Decomposition for Figure of Merit Statistical Errors . . 31

2.2.8 Spatial Statistics Analysis of Chemical Datasets 32

2.2.8.1 Basic Categories of Dataset Topology 32

2.2.8.2 Refined Nearest Neighbor Analysis: Mathematical Foundations 34

2.2.8.3 Refined Nearest Neighbor Analysis: Implementation . . 36

2.2.9 Visualization of Topology by Self-Organizing Maps (SOMs) . . . 39

2.2.10 Measures of Correlation 41

2.3 Results 43

2.3.1 Characterization of Benchmark Dataset Sub-Samples by Refined Nearest-Neighbor Analysis 43

2.3.2 Correlation of VS Performance and Dataset Clumping 43

2.3.3 Topology Induced Component of Variance 48

2.3.4 Comparison of Refined Nearest Neighbor Analysis with Other Approaches for Dataset Analysis 50

2.3.5 Mapping Performance 52

2.3.6 Application to Whole Datasets 54

2.3.7 Benchmark Dataset Bias 58

2.4 Summary 60

3 Maximum Unbiased Validation (MUV) Datasets **62**

3.1 Objectives 62

3.2 Methods 63

3.2.1 Criteria for Refined Nearest Neighbor Analysis Based Benchmark Dataset Design 63

3.2.2 MUV Benchmark Dataset Design Strategy 64

3.2.3 PubChem as a Source of VS Validation Datasets 65

3.2.4 Selection of Bioactivity Datasets 67

3.2.5 Assay Artifacts Filter 67

3.2.5.1 Hill Slope Filter 69

3.2.5.2 Frequency of Hits (FoH) Filter 69

3.2.5.3 Autofluorescence and Luciferase Inhibition Filter . . . 73

3.2.6 Potential False Negatives in the Datasets of Decoys 74

3.2.7 Chemical Space Embedding Filter 75

3.2.8 Descriptors 77

3.2.9 Spatial Statistics for Benchmark Dataset Design 77

3.2.9.1 Preliminary Experiments for the Determination of t_i . . 77

3.2.9.2 Spatial Statistics for PubChem Datasets 81

3.2.10 Design of MUV Datasets 81

3.2.11 Retrospective virtual screening simulations 84

3.2.12 Figures of Merit (FoM) for Virtual Screening Performance 84

3.2.13 Unique Molecular Frameworks 86

3.3 Results and Discussion 88

3.3.1 Bioactivity Datasets Extracted from PubChem 88

3.3.2 MUV Benchmark Datasets: General Properties 88

3.3.3 Spatial Statistics Analysis of MUV Datasets 90

3.3.4 Application of MUV Datasets for LBVS Benchmarking 93

3.3.5 Comparison of MUV with DUD 97

3.4 Summary 101

4 Conclusion and Outlook **103**

A Supplementary Tables: Chapter 2 **106**

B Supplementary Tables: Chapter 3 **130**

C MUV Datasets **143**

D Spatial Statistics Toolbox for MATLAB 7 **145**

D.1 Installation 145

D.2 Source Code 147

D.2.1 $G(t)$ - spst_G 147

D.2.2 $F(t)$ - spst_F 148

D.2.3 Design of Datasets of Actives with Given ΣG - spst_RX 149

D.2.4 Design of Datasets of Decoys with Given ΣF - spst_GA 151

D.3 How To Generate Your Own MUV Datasets 156

D.3.1 Basic Preparations 156

D.3.2 Calculate Simple Descriptors 158

D.3.2.1 Calculating Simple Descriptors using OpenEye FILTER and BABEL 159

D.3.2.2 Calculating Simple Descriptors Using the CDK 161

D.3.3 Generating Maximally Diverse Sets of Actives and Minimally Separated Sets of Decoys 161

D.3.4 Generating MUV Datasets 162

D.3.5 Exporting the Datasets to SD-Files 163

Bibliography **165**

Symbols and Abbreviations **183**

Index **187**

Curriculum Vitae **191**

Abstract

A common finding of many reports evaluating ligand-based virtual screening methods is that validation results vary considerably with changing benchmark datasets. It is widely assumed that these dataset specific effects are caused by the redundancy, self-similarity and cluster structure inherent to those datasets. These phenomena manifest themselves in the datasets' representation in descriptor space, which is termed the dataset topology. A methodology for the characterization of dataset topology based on spatial statistics is introduced. The method is non-parametric and can deal with arbitrary distributions of descriptor values. With this methodology it is possible to associate differences in virtual screening performance on different datasets with differences in dataset topology. Moreover, the better virtual screening performance of certain descriptors can be explained by their ability of representing the benchmark datasets by a more favorable topology. It is shown, that the composition of some benchmark datasets causes topologies that lead to over-optimistic validation results even in very "simple" descriptor spaces. Spatial statistics analysis as proposed here facilitates the detection of such biased datasets and provides a tool for the design of unbiased benchmark datasets.

Based on the aforementioned results, general principles for the design of benchmark datasets, which are not affected by topological bias, were developed. Refined Nearest Neighbor Analysis was used to design benchmark datasets based on PubChem bioactivity data. A workflow is devised that purges datasets of compounds active against pharmaceutically relevant targets from unselective hits. Topological optimization using experimental design strategies monitored by Refined Nearest Neighbor Analysis functions was applied to generate corresponding datasets of actives and decoys that are unbiased with regard to analogue bias and artificial enrichment. These datasets provide a tool for an Maximum Unbiased Validation (MUV) of virtual screening methods. The datasets and a MATLAB toolbox for spatial statistics are freely available on the enclosed CD-ROM or via the internet at http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html.

Zusammenfassung

Ein gemeinsames Ergebnis vieler Arbeiten zur Validierung von Methoden des ligandbasierten Virtuellen Screenings ist die Tatsache, dass die erzielten Ergebnisse stark von den verwendeten Validierdatensätzen abhängen. Es wird angenommen, dass diese datensatzspezifischen Effekte durch die Redundanz, Selbstähnlichkeit und Clusterstruktur verursacht werden, die in vielen dieser Datensätze auftreten. Die Abbildung eines Datensatzes im Deskriptorraum, die "Datensatztopologie" (engl. *Dataset Topology*), spiegelt diese Phänomene wider. Im Rahmen der Arbeit wird eine nicht-parametrische Methode aus dem Bereich der räumlichen Statistik (engl. *Spatial Statistics*) zur Charakterisierung der Datensatztopologie eingeführt. Mit dieser Methode ist es möglich, Unterschiede in den Ergebnissen von Validierexperimenten mit Unterschieden in der Datensatztopologie zu erklären. Darüberhinaus kann das bessere Abschneiden einiger Deskriptoren mit deren Fähigkeit erklärt werden, günstigere Topologien im jeweiligen Deskriptorraum zu erzeugen. Die Zusammensetzung mancher Validierdatensätze bedingt Topologien, die zu überoptimistischen Validierergebnissen führen. Die vorgestellte Methodik aus dem Bereich der räumlichen Statistik ermöglicht es, solche Datensätze vor der Validierung zu erkennen. Weiterhin kann die Methode verwendet werden, um zielgerichtet Datensätze zu konstruieren, die unverfälschte, nicht von der Datensatzzusammensetzung beeinflusste Validierergebnisse sicherstellen.

Auf diesen Ergebnissen aufbauend werden generelle Kriterien für die Konstruktion von Validierdatensätzen ohne topologiebedingte Verzerrung entwickelt. Mit Hilfe von Methoden der "Verfeinerten Nächster Nachbar Analyse" (engl. *Refined Nearest Neighbor Analysis*) werden verzerrungsfreie Datesätze zur Validierung von Techniken des Virtuellen Screenings generiert. Als Basis dienen Datensätze von Substanzen mit Bioaktivität gegen pharmazeutisch relevante Zielproteine aus PubChem. Ein im Rahmen der vorliegenden Arbeit neu entwickeltes Verfahren ermöglicht es, Substanzen mit unspezifischer Bioaktivität aus diesen Datensätzen zu entfernen. Durch Optimierung der Datensatztopolo-

gie werden korrespondierende Datensätze von Aktiven und Inaktiven erstellt, die eine Maximal Unverfälschte Validierung (MUV, engl. *Maximum Unbiased Validation*) von Techniken des Virtuellen Screenings ermöglichen. Diese Datensätze und eine MATLAB Toolbox für räumliche Statistik sind auf der beiliegenden CD-ROM oder im Internet unter http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html frei verfügbar.

Chapter 1

Introduction

1.1 Virtual Screening in the Process of Drug Discovery

1.1.1 Screening: An Entry-Point to Drug Discovery

The initial search for chemical compounds that exert a pharmacological effect is an important step in modern pharmaceutical research. Such *lead* compounds are the starting point of medicinal chemistry programs that produce large series of derivatives of the lead compound in order to optimize its pharmacological and biological properties. Traditionally, lead compounds were found - if not by serendipity - by the isolation of active compounds from e.g. plants, fungi, microorganisms or animal poisons.[1] Although the purification, synthesis or modification of natural products has lead to some major successes of pharmaceutical research, the difficulties in synthesizing and purifying more complex natural products and their derivatives have proven to be a major bottleneck.[1,2]

In 1904, Paul Ehrlich and Kiyoshi Shiga conducted a ground-breaking experiment, when they tested a chemical library of 100 azo dyes for their effect on mice infected with Trypanosoma.[2,3] This research resulted in the discovery of Nagana Red, the first synthetic drug against bovine nagana disease. In the wake of this success, an extensive medicinal chemistry program was conducted including the testing of thousands of derivatives and resulting in the discovery of Suramin in 1916,[2] which is still in use today as one of the major therapeutics against Trypanosomiasis.[4] Nowadays, this *screening* of chemical

libraries is established as the first step in drug discovery research. However, in contrast to the days of Paul Ehrlich, screening is usually no longer conducted using animals as test systems.

Today, before the start of a drug discovery campaign, a molecular target of pharmaceutical relevance is usually identified by molecular biology methods.[5] Knowledge about the biochemistry, structure or cellular biology of this target is then utilized to screen large libraries often consisting of several hundreds of thousands of compounds. In order to cope with the magnitude of this task, two major techniques of screening for promising lead compounds have been established in recent years: Experimental High-Throughput Screening (HTS) and computer based Virtual Screening (VS), which are also often referred to as *in vitro* and *in silico* screening, respectively.

1.1.2 Experimental High-Throughput Screening (HTS)

Most pharmaceutical companies employ a series of miniaturized and automated *in vitro* tests as the first step in lead identification, once a biological target has been identified and validated.[6] Using such assays, it is possible to screen more than 100.000 chemical compounds against the isolated protein target or a specialized cell system in a reasonable amount of time. Such experimental High-Throughput Screening (HTS) has become a routine step in every drug discovery program of large pharmaceutical companies.[6]

Basically, HTS is an automated process, that rapidly assays large numbers of compounds against a biological target. (Figure 1.1) Robotic systems are employed to automatically subject each compound in a large collection of chemicals to a miniaturized biological assay. Usually, only a single standard concentration of each compound is tested in these *primary* screening experiments. Since this approach is prone to a relatively high occurrence of errors,[7–9] compounds showing promising results in the primary screen are later re-assessed using classical low-throughput techniques in a so-called *confirmation* assay. The assay technologies used in modern HTS can be traced back to three different origins, which prevail today for a rough classification of assay formats and functions.[10]

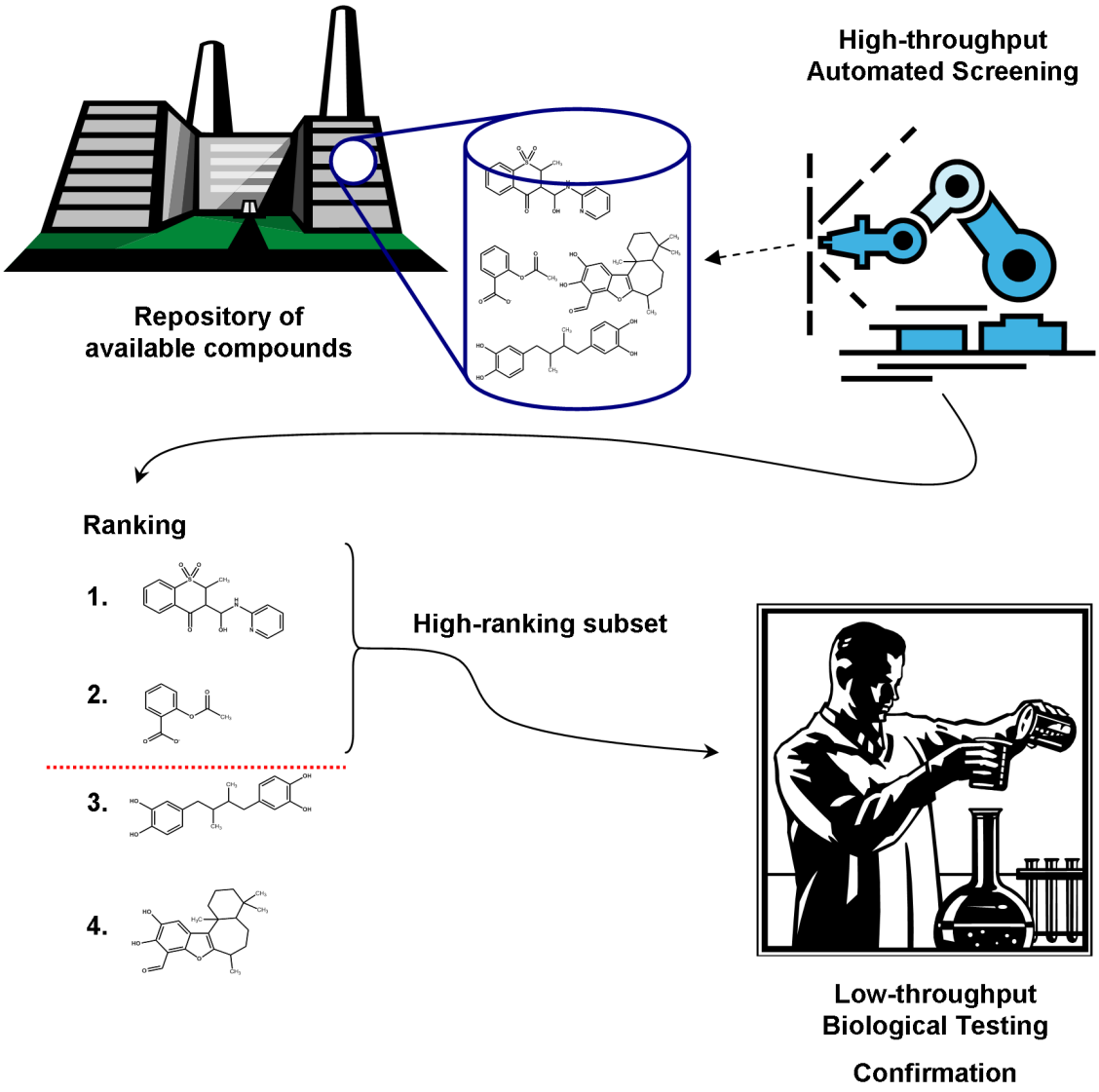


Figure 1.1: Typical High-Throughput Screening (HTS) workflow. A large number of compounds from a repository of chemical samples is screened using fast, automatized biological assays. Compounds are ranked according to their measured biological activity. Since automated HTS has a relatively high error rate,[7–9] a top-ranking subset is chosen and re-examined using classical low-throughput biological assays.

1.1.2.1 Common HTS Assay Formats

Cellular Growth and Proliferation Assays. Assays measuring the inhibition of cellular growth or proliferation have a long tradition of application in the search for leads in the fields of anti-infectives or cancer.[10] They simply measure if a certain chemical is able to reduce or inhibit the growth of a test population of cells. Growth and proliferation assays are prototypes of so-called "phenotypic" assays, meaning that the assignment of activity or inactivity is solely based on the "phenotypic" observation of reduced growth, without any knowledge about the underlying molecular target or mechanism. Indeed, it is impossible to identify the molecular target of active compounds in many cases. The simple technical realization of growth and proliferation assays is therefore often balanced by the difficulty of discriminating pharmacologically meaningful results from cytotoxic effects.[10]

Functional Cell-based Assays. Classical pharmacological bioassays frequently relied on measurements of the contractile state of smooth muscle cells. It was realized early, that sensitivity and specificity towards a specific kind of cell-surface receptor were essential for the bioassay to generate meaningful results.[10] Recent advances in molecular biology provided cell-lines specifically expressing drug target proteins as cell-surface receptors. Equally specific down-stream effects, such as Ca^{2+} release or reporter gene activity provide experimental readouts amenable to a high-throughput approach. Although functional cell-based assays involve a phenotypic observation and therefore share several of the shortcomings of growth and proliferation assays, the observed effect is usually highly specific for an interaction with the examined target. A special advantage of functional cell-based assays is the possibility to test for all types of modulation, i.e. inhibition, activation and allosteric interactions in one assay.

Enzymatic Assays. Because of recent advances in molecular biology, biochemistry and biotechnology, many target proteins can be recombinantly expressed and purified in sufficient amounts to facilitate high throughput screening using isolated enzymes. A great variety of enzymatic binding assays based on competition with labeled ligands or substrates

is available. It is evident, that enzymatic assays ensure the highest degree of specificity for the observed effect, while providing the experimenter with a well defined system in which different sets of parameters can easily be tested.[10] However, it is often difficult or impossible to detect activation or allosteric modulation in competitive binding assays. In contrast to cell-based assays, enzymatic assays also exclude any effects exerted by cellular membranes or cellular compartmentalization from the test. However, such effects might be of considerable pharmaceutical relevance in later development stages. Moreover, enzymatic assays are highly sensitive towards the tendency of certain chemical compounds to form aggregates when diluted in small volumes of liquid as they are common on HTS micro-plates.[11]

1.1.2.2 PubChem: HTS Data for the Public Sector

HTS is an established approach in the pharmaceutical industry and is considered one of the main resources for the discovery of new drugs.[10] However, due to its high costs, the public and academic sector has not been able to make extensive use of the technology. Therefore, the Molecular Libraries Initiative (MLI)[12] was devised as part of the NIH Roadmap for Medical Research[13,14] to provide academic researchers with results from HTS campaigns conducted by a consortium of public sector screening facilities, the Molecular Libraries Screening Centers Network (MLSCN). The screening data produced by the MLSCN laboratories is stored centrally in a data repository called PubChem,[15] which is publicly available via the internet and can be automatically interfaced by web capable computer programs.[16] PubChem is composed of three major databases: (i) PCCompound provides chemical structures of the compounds tested as part of the NIH Roadmap effort. (ii) PCBioAssay lists bioactivity data for presently (July 2008) more than 640,000 compounds, derived from readouts of more than 1100 bio-assays, which include both, primary and confirmation screens. (iii) PCSubstance contains data such as supplier information, registration IDs or links to other databases for the actual physical samples of substances tested in the HTS assays. If the contents of these samples are chemically characterized, PCSubstance also provides links to the respective structures in PCCompound.

The fact that all data in PubChem is public and easily accessible by computer programs, makes PubChem an invaluable tool for chemoinformatics analyses of bioactivity data of small molecules. It will therefore serve as one of the main data resources of this study. (see Chapter 3)

1.1.3 Virtual Screening (VS)

Although experimental High-Throughput Screening has generated considerable success and consequently is one of the mainstays of modern drug discovery, its considerable costs and the storing capacities needed for ever-growing compound libraries have put limitations to its use.[17] Therefore, computer based methods such as Virtual Screening (VS) are today applied as a standard technique in almost every drug discovery campaign, both in industrial and academic environments. Their main goal is to narrow down large databases of chemical compounds to a level, where experimental scientists or automated systems can cope with testing the substances for biological activity.[6] That way the number of compounds actually subjected to costly biological testing procedures is greatly reduced. (Figure 1.2)

Virtual screening methods can be roughly divided into structure-based and ligand-based approaches.[18,19] This study will focus on ligand based virtual screening (LBVS). However, both approaches will be introduced briefly using two well-known inhibitors of the target protein Cyclooxygenase 2 (COX2), Celecoxib and Rofecoxib (Figure 1.3), as examples.

1.1.3.1 Structure Based Virtual Screening (SBVS)

The basic rationale of structure based virtual screening (SBVS) approaches is Fischer's "Lock and Key" principle, a central paradigm in biological chemistry and medicinal chemistry.[20] According to Fischer's principle, the exertion of a pharmacological effect by a chemical compound on a target protein requires the formation of a specific, energetically favorable, three dimensional pattern of chemical interactions.[20] The goal of SBVS is to predict for each compound in a virtual library of potential leads or drugs, if and to what

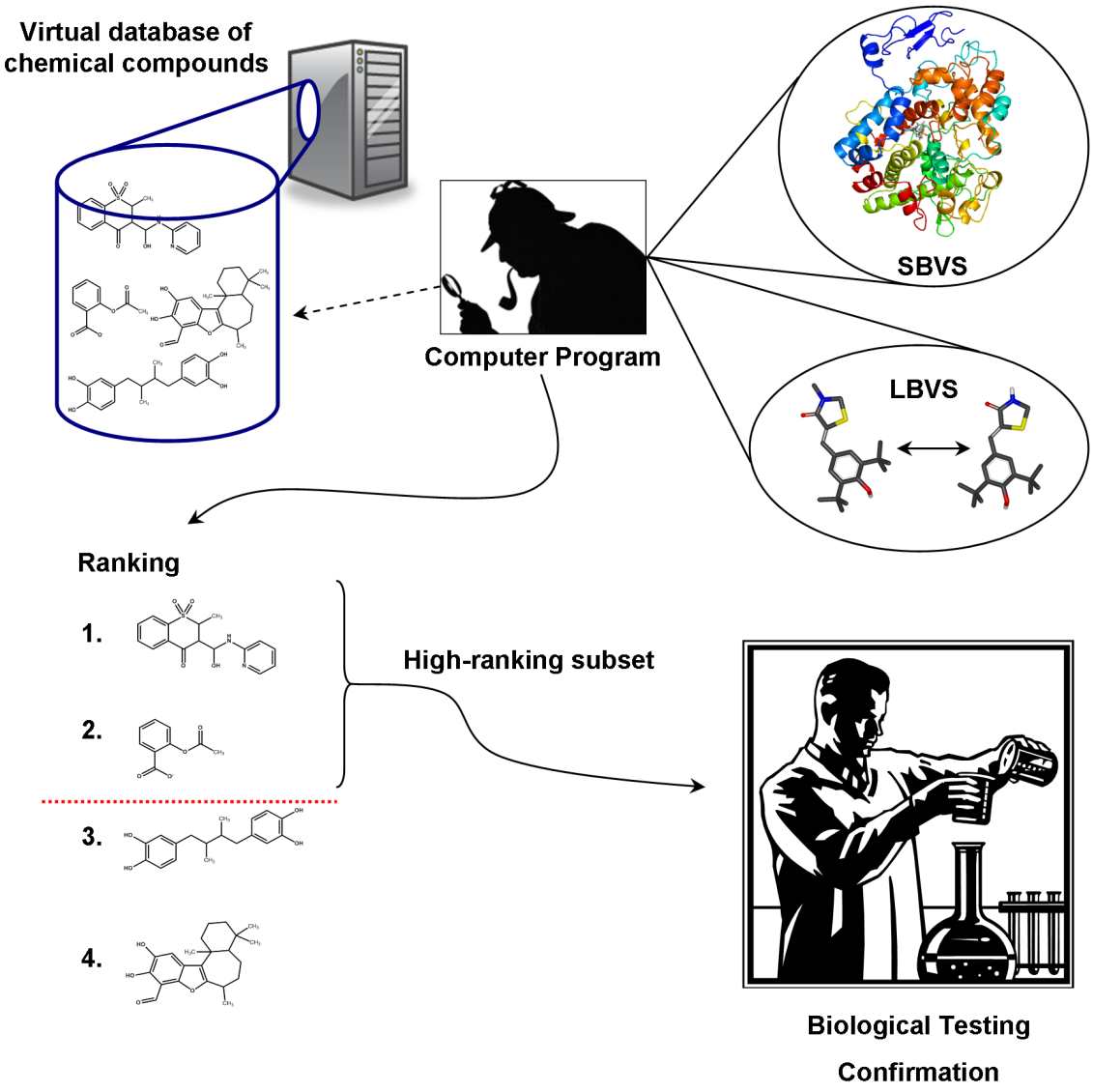
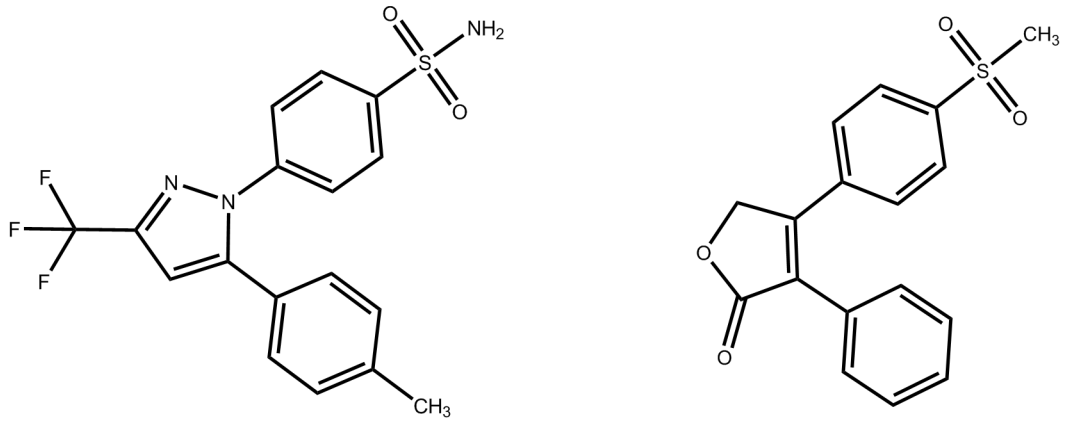


Figure 1.2: Virtual Screening (VS) for lead compounds. A great advantage of VS is the fact, that the compound library needed for screening is virtual, i.e. the screened compounds do not need to be present in physically. The virtual library is screened by a computer program applying either structure based virtual screening (SBVS, see Section 1.1.3.1) or ligand based virtual screening (LBVS, see Section 1.1.3.2) techniques. The result is a rank ordered list of the molecules in the database, sorted according to their predicted activity. From this list, a subset of top-ranking compounds is chosen for biological testing. Provided the top-ranking compounds are enriched in active compounds, the amount of tested compounds can be greatly reduced, thereby shortcutting the first iteration of experiments in the HTS workflow. (compare Figure 1.1)



Celecoxib **Rofecoxib**

Figure 1.3: COX2 inhibitors: Celecoxib and Rofecoxib.

extent it can interact energetically favorably with a given target protein. Therefore, a prerequisite for the conduction of SBVS is the availability of a three-dimensional structure of the respective protein target either from experiments or computational predictions.[18] If this condition is met, "molecular docking", i.e. the prediction of complexes of the target and a compound, can be conducted in a two step process. "Posing" (Figure 1.4) constitutes the enumeration of an ensemble of possible binding modes between compound and target. "Scoring" assigns a score, that is proportional to the predicted binding energy, to each of the generated poses.[21] The basic assumption of SBVS is that compounds with large negative, i.e. favorable, predicted binding energy, bind the target specifically and might exert a pharmacological effect. Following this assumption, compounds are classified as presumably active or inactive against the target by the score of their best docking pose.[21]

1.1.3.2 Ligand Based Virtual Screening (LBVS)

In contrast to SBVS approaches, ligand based virtual screening (LBVS) does not require any knowledge about the three-dimensional structure of the target protein. This is of considerable advantage in the quest for binders of targets or target classes, for which it is difficult to obtain structural data. Some of these target classes, such as for instance G-protein coupled receptors (GPCR)[26] or ion channels,[27] are of considerable pharma-

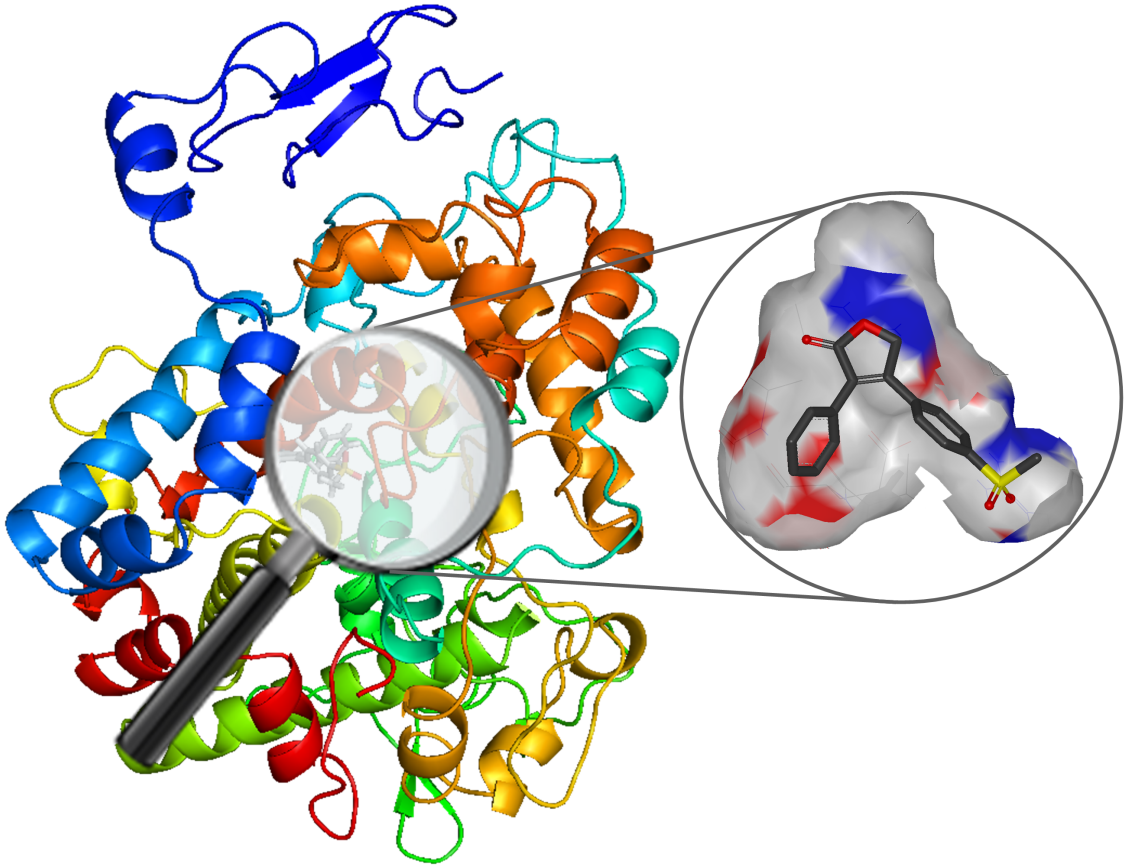


Figure 1.4: One possible binding mode (pose) of Rofecoxib in Cyclooxygenase 2 (COX2) as predicted by the docking program FRED2.[22] Knowledge about the three-dimensional structure of COX2 is essential to derive the geometry and contours of the binding pocket (Inset). Steric complementarity between the coxib scaffold and the binding pocket is observable. Binding of the methyl sulfonyl moiety to the side pocket of the binding site (Inset: lower right) increases COX2 selectivity.[23] *Docking: OpenEye FRED2,*[22] *Visualization: PyMol,*[24] *OpenEye VIDA3*[25]

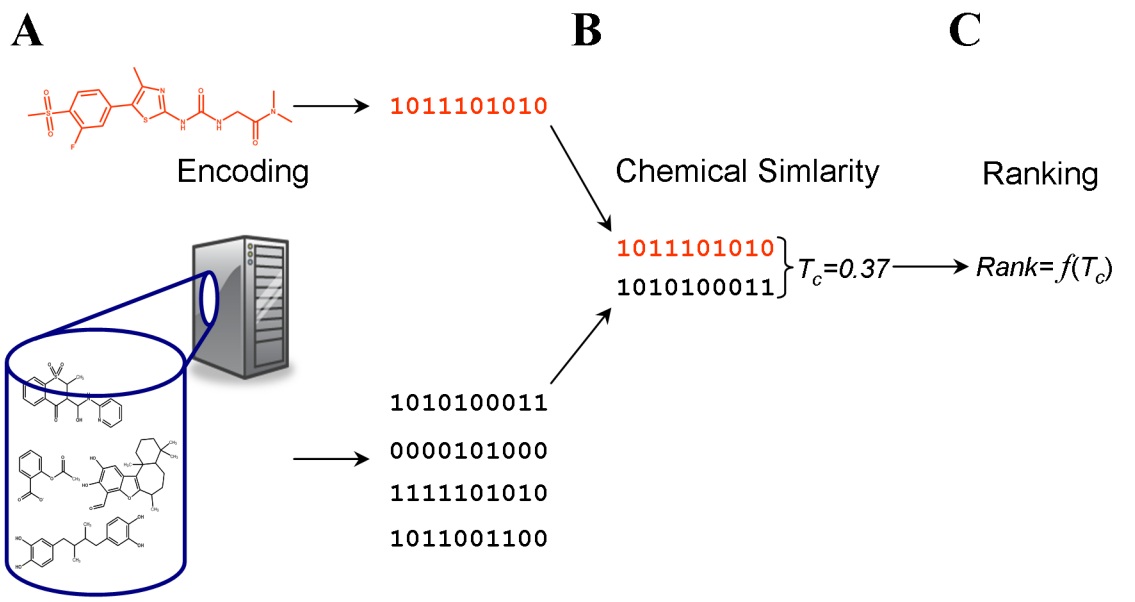


Figure 1.5: Overview of the ligand based virtual screening (LBVS) process. (A) Query and database compounds are encoded by pharmacophore patterns or descriptor vectors. (B) Chemical similarity between query and database compounds is quantified. (C) The library compounds are ranked according to their similarity with the query.

ceutical relevance. The omission of structural information is possible, because LBVS relies on another fundamental paradigm of medicinal chemistry that is independent of the target protein. This is the so-called similarity principle, which states that chemically similar substances are likely to share similar biological activity.[28–31] As a consequence, the basic precondition for the conduction of LBVS is the knowledge about one or several compounds active against the target under scrutiny. These compounds are often called the "query". The goal of LBVS is to rank a database of potential leads or drugs according to their similarity with the query, following the assumption that compounds, which are very similar to the query, are likely to exhibit the same, i.e. the desired, biological activity. LBVS involves three basic steps (Figure 1.5):

Encoding of compounds. The physical and chemical properties of the molecules in the query and the database to be screened must be represented numerically so that they can be used as input for a mathematical quantification of chemical similarity. A large variety of approaches exists for this encoding process,[19] which can be roughly divided into pharmacophore methods (Figure 1.6) and descriptor methods. (Figure 1.7)

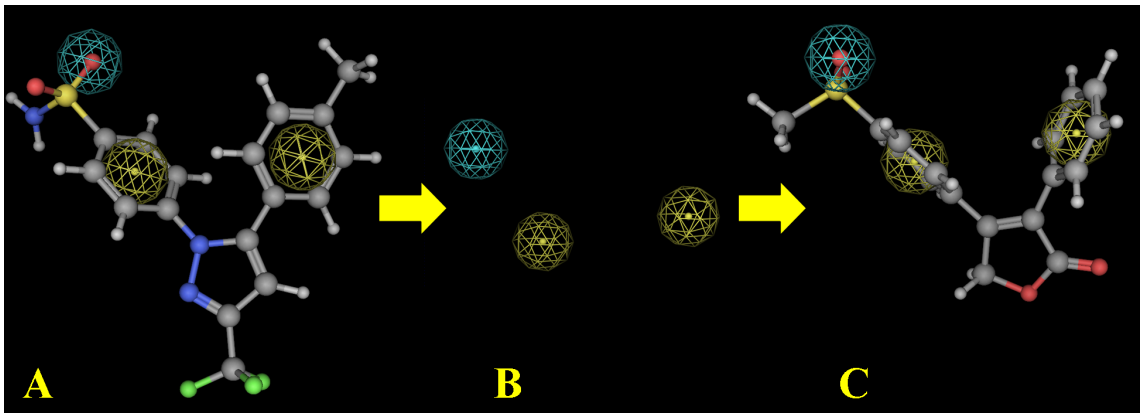


Figure 1.6: Pharmacophore search using Celecoxib as a reference structure. (A) From the three-dimensional structure of Celecoxib, a geometric pattern of two aromatic rings (yellow spheres) and a H-Bond acceptor (blue sphere) can be derived as one possible element crucial for COX2 inhibition. (B) This pharmacophore can be used to search a database of potential drugs. (C) Rofecoxib matches the requirements of the pharmacophore query and is identified as a COX2 inhibitor. *Visualization: MOE*[34]

A pharmacophore is a specific geometric arrangement of a set of structural features in a molecule that is recognized at a receptor site and is responsible for that molecule's biological activity.[32,33] Pharmacophore queries can be derived from one or several active compounds either manually or algorithmically. (Figure 1.6)

Descriptors are vectors calculated from the chemical structure of a compound. A large variety of such descriptors with different levels of sophistication and complexity has been developed.[19] However, among the literally thousands of available descriptors[35] three coarsely grouped classes prevail: (i) substructure binary fingerprints, where each bit codes for the absence or presence of a distinct substructural feature in the molecule, (ii) topological and geometric descriptors based on mutual distances between molecular features and (ii) molecular property vectors formed of the numerical values of physico-chemical properties like the logarithm of the octanol/water partition coefficient (logP) or the molecular weight.[19] (Figure 1.7) In addition to "pure" implementations of the concepts of pharmacophores and descriptors, combined approaches exist, in which the presence, absence or abundance of certain pharmacophoric patterns in a molecule is used to calculate a fingerprint or descriptor. These approaches have proven especially useful, since they combine the biochemical relevance of the pharmacophore concept with the mathematical versatility of descriptor vectors.[36–39]

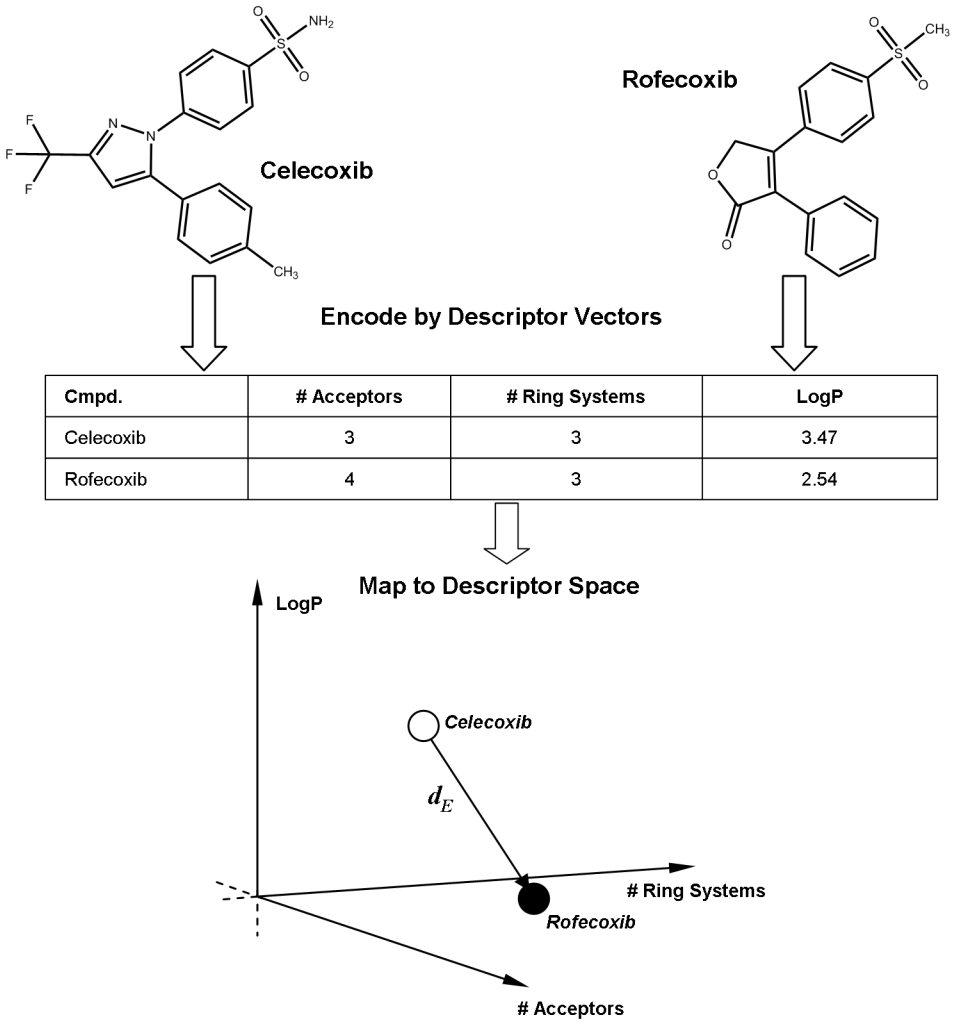


Figure 1.7: Molecular features such as the number of H-Bond acceptors and the number of ring systems in a molecule, or physico-chemical properties like LogP can be calculated from the molecular structure. If they are utilized to form the components of descriptor vectors, these vectors provide a means for the mathematical quantification of molecular similarity by mapping chemical compounds to a vector space. Vicinity in this descriptor space is equivalent to chemical similarity, which can be measured by geometric distance measures such as the Euclidean distance d_E.

Quantification of chemical similarity. Based on the numerical representation by pharmacophores or descriptors, the chemical similarity between query and database compounds needs to be quantified. Here, descriptors have the unique advantage of being vectors. This allows the use of a large variety of distance measures that are well established in analytical geometry.[29] The most widely used are the Euclidean distance for continuous descriptor vectors and the Tanimoto-Jaquard coefficient for bitstring fingerprint vectors.[40,41]

The Euclidean distance is given as:

$$d_E = \sqrt{\sum_i (x_i - y_i)^2}; \tag{1.1}$$

where x_i, y_i are the i^{th} components of the descriptor vectors **x** and **y** of two molecules.

The Tanimoto-Jaquard coefficient is given as:

$$T_c = \frac{n_{xy}}{n_x + n_y}; \tag{1.2}$$

where n_{xy} is the number of bits set "on" in the vectors of both molecules and n_x, n_y the number of bits set "on" in the vectors **x** or **y**, respectively.

In addition to these important measures for descriptor similarity a large variety of other distance functions exist, that can map all levels and nuances of molecular similarity.[29] In this context, spatial vicinity in the vector space spanned by the respective descriptor - often called "descriptor space" - is equivalent to molecular similarity.[42] (Figure 1.7) In contrast to this, classical pharmacophore searches can only determine if a given molecule does or does not match a certain pharmacophoric pattern.

Database ranking Given a molecular query and a database encoded by pharmacophores or descriptors and a method for the quantification of molecular similarity, an algorithm must be applied to generate a ranking of the compounds most similar to the query. A wide range of search methodologies are available to accomplish this task.[19,29] The most basic approach is to rank-order the screening database compounds according to their similarity to a single query molecule. Aside from these most basic "single query" similarity

searches, methods featuring all levels of complexity exist, including for instance consensus methods,[43] artificial neural networks (ANN)[44] or support vector machines (SVM).[45] Recent results have also demonstrated the usefulness of quantitative structure activity relationship (QSAR) models in LBVS applications.[46] A particularly efficient method for LBVS with multiple query molecules is similarity searching using data fusion according to the "MAX"-rule.[47,48] The similarity of each molecule in the screened database with each molecule in the query is calculated, and the maximum of these values of similarity, i.e. the nearest neighbor similarity, determines the rank of the molecule in the search output. This approach has proven to be extremely powerful both in our own experience with the conduction of virtual screenings and in comparative studies in the literature.[47–49]

The focus of this study will be descriptor based techniques for ligand based virtual screening. However, the presented results and methods are readily applicable to pharmacophore searches and require only minor modifications for the application to SBVS methods.

1.2 Validation of Virtual Screening Techniques

1.2.1 Objectives of Validation Experiments

The utilization of similarity in descriptor space as a predictor of biological activity poses a question of obvious importance: how much chemical similarity is enough to ensure similar biological activity? Put more mathematically: An estimate is required for the numerical cut-off value of similarity that is sufficient to predict similar biological activity with minimum error.

In a seminal paper, Martin et al. have shown that it is not possible to determine an explicit numerical similarity-cutoff in descriptor space that ensures common biological activity in the general case.[30] However, in two earlier papers Brown and Martin assessed the ability of various descriptors to discriminate actives from inactives[50] and to predict physico-chemical properties relevant to receptor binding[51] using datasets of molecules for which these parameters were experimentally determined. This benchmark dataset based

validation methodology has become a standard procedure for the evaluation of chemical similarity methods, structural descriptors and search algorithms, because it estimates a parameter deemed critical for prospective virtual screening campaigns: the number of hits to expect for a particular method. There are typically two types of VS validation experiments: (i) In "benchmarking" experiments, the objective is to identify the method with the best VS performance across a range of datasets. Experiments of this kind are most often applied to show how well a novel VS method performs in relation with available ones[52–54] or by pharmaceutical companies in order to determine which software to acquire.[55,56] (ii) "Suitability testing" experiments are employed to determine the method or set of operational parameters of a method best suited for a particular target or target class. Results from suitability testing experiments may vary considerably depending on the dataset of active compounds and their respective target. Experiments of this type are usually employed in the preparation of prospective VS campaigns in order to facilitate a rational choice of the optimal method for the target under investigation. Although the final results and conclusions obtained from suitability testing and benchmarking are not comparable, two basic tasks for VS validation can be formulated, that are valid under both experimental settings:

(1) Compare the performance of different virtual screening methods.

(2) Estimate the number of hits a method is likely to retrieve from a database.

1.2.2 Validation Procedures and Figures of Merit (FoM)

For both, benchmarking and suitability testing, a validation procedure normally starts with the selection of a set of molecules with known activity against the target under scrutiny. Part of this set is chosen (often randomly) to act as query. The rest is pooled with a usually large number of inactive molecules (frequently called "decoys") to become the validation set. The VS method to be validated is then applied to the validation set and assessed based on the produced ranking. For descriptor based LBVS methods, both sets, the query and the validation set, are encoded by the descriptor to be validated and the validation set is ranked according to its similarity with the query. (Figure 1.8)

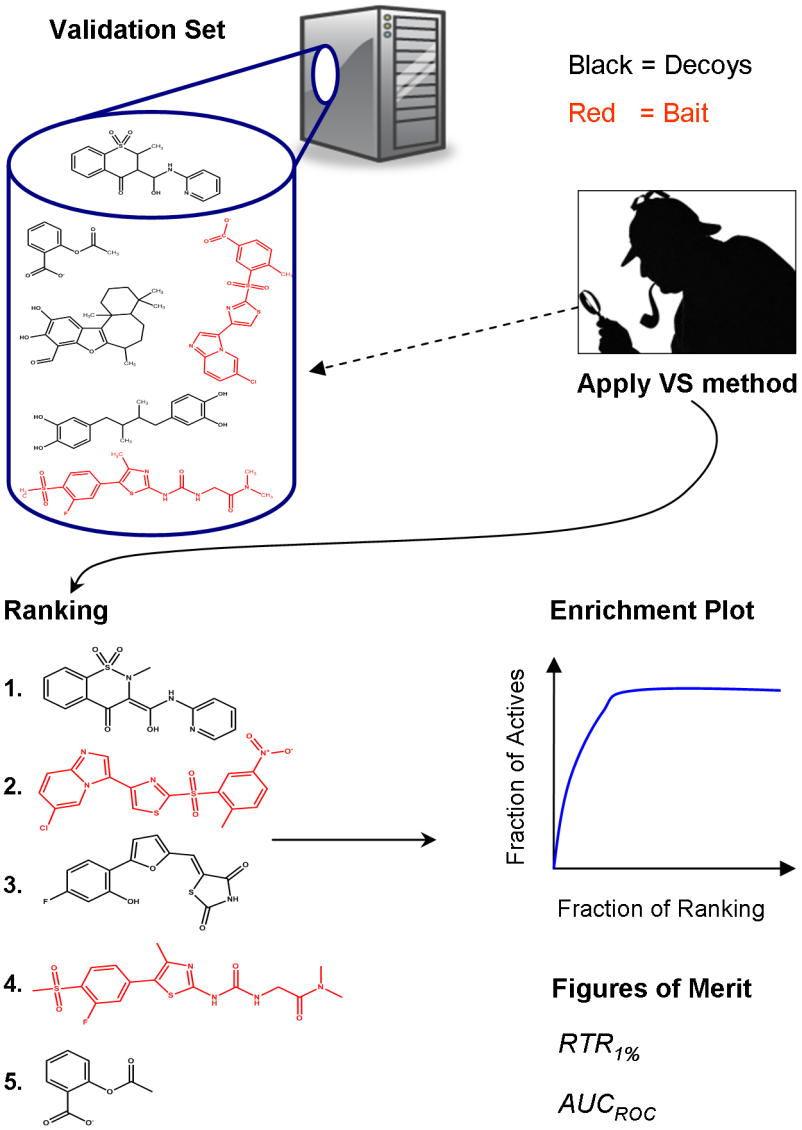


Figure 1.8: Typical validation process. A database of inactive compounds (decoys, black) is spiked with a set of known actives (red) to become the validation set. The VS method to be validated is applied to the validation set to produce a ranking. This ranking is used to assess the performance of a VS method. Plots of the fraction of found actives vs. the fraction of the database screened, so-called enrichment plots, are used to assess the quality of a method. These plots are also the basis for the calculation of Figures of Merit, such as $RTR_{1\%}$ and AUC_{ROC}.

A common way of visualizing the retrieval of known active substances in the resulting VS ranking is to plot the fraction of found actives against the fraction of the ranking containing it. These plots are generally called "enrichment plots". (Figure 1.9A) The recall or retrieval rate (RTR) at one percent of the ranked validation set ($RTR_{1\%}$) is the fraction of active compounds that is retrieved in the first percent of the ranking generated by the validation run. It can be read out from an enrichment plot at $x = 0.01$ ($= 1\%$). (Figure 1.9B) The $RTR_{1\%}$ has been established as a widely used figure of merit for the performance of VS methods.[47,48] It is evident, that the RTR can be used to compare methods and to estimate the expected number of hits. Thereby it fulfills both basic tasks of VS validation as stated above. Often so-called "enrichment factors" (EF)[57] are calculated from the RTR, that are meant to normalize to the null hypothesis of uniformly distributed active molecules in the final ranking list. Bender and Glen have shown that enrichment factors tend to overestimate performance, since EFs quantify the performance of a method relative to the hypothesis of uniformly distributed actives in the final ranking, which is not realistic.[58] Moreover, several authors have criticized the RTR and EF metrics for the fact of being susceptible to changes in the ratio of the sizes of benchmark dataset vs. background and the inability to reflect the position of the found actives before the threshold.[59,60]

In order to avoid these issues, the area under the receiver operating characteristic curve (AUC_{ROC}, Figure 1.9D) was used in a number of studies for the analysis of VS validation rankings.[61–64] In receiver operating characteristic curves (ROC), the true positive rate (the fraction of found actives) is plotted against the false positive rate (the fraction of found decoys). (Figure 1.9C) The AUC_{ROC} metric has one important shortcoming: it is unable to address the so-called "early recognition" problem. Since usually only a small fraction of a VS ranking can be tested experimentally, a good metric for VS should reflect the enrichment of actives at the beginning of the ranking. In addition, *ROC* metrics do not provide a direct estimate of the number of expected hits. Recent efforts have sought to develop VS performance metrics that combine the statistic stability of the AUC_{ROC} with the "early recognition" properties of the RTR or the EF.[59,60] However, due to their novelty, these metrics have not yet found extensive use in VS validation studies and it is therefore

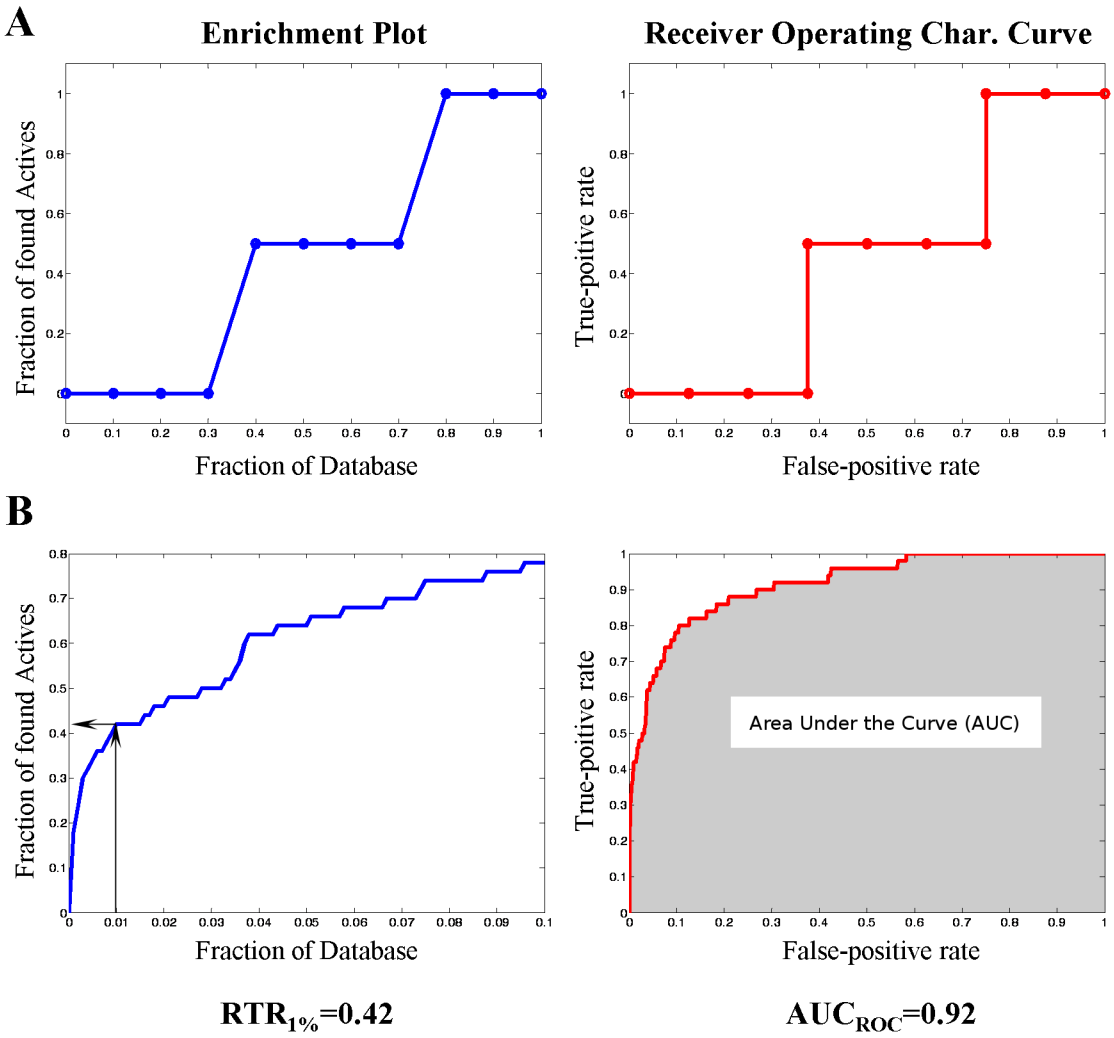


Figure 1.9: (A) Enrichment plot (blue) and receiver operating characteristic curve (red) for an artificial ranking in which two out of a total of two actives are found at rank $r = 4,8$ in a database of 10 compounds. (B) Enrichment plot (blue) and receiver operating characteristic curve of a ranking obtained from an actual VS validation with 50 actives among roughly 90000 decoys. Both, $RTR_{1\%} = 0.42$ and $AUC_{ROC} = 0.92$ indicate quite satisfactory validation results.

difficult to compare the results obtained with these metrics to those of other works. Thus, in this study, $RTR_{1\%}$ and AUC_{ROC} were used in a complementary manner for the analysis of VS rankings.

1.3 Benchmark Datasets

1.3.1 Available Benchmark Datasets for VS Validation

A basic condition for the conduction of validation experiments as proposed by Brown and Martin[50,51] is the availability of commonly employed benchmark datasets. A number of such benchmark datasets have been published for both, ligand based virtual screening and structure based virtual screening .[47,48,56,64–67] They range in size from tens to several hundreds of active compounds. The sources of these datasets vary, including the medicinal chemistry literature, public databases like the PDB,[68] commercial databases like the MDL Drug Data Report (MDDR)[69] or proprietary data of pharmaceutical companies. As a consequence, the experimental conditions leading to the qualification of a compound as active usually vary within each dataset. It is often tedious, in the case of literature datasets, or impossible, in the case of proprietary data, to examine and compare these experimental conditions. In addition to these difficulties regarding the datasets of actives, many literature benchmark datasets share a significant shortcoming: the inactivity of the decoy compounds is not experimentally determined but merely assumed.[47,48,56,64–67,70] The decoys are usually extracted from large databases of drug-like compounds, such as ZINC[71] or the MDL Drug Data Report (MDDR)[69], simply by using compounds with no reported activity against the target under scrutiny.[47,48,56,64–67,70]

However, regardless of the shortcomings they may have, benchmark datasets are vital for a rational evaluation of VS methods. Especially the datasets referenced above have proven to be highly valuable to the field of VS validation studies. From these, the datasets provided by Hert and Willet in two seminal papers[47,48] stand out as the ones most widely accepted for the validation of virtual screening methods. They comprise 11 datasets, each consisting of several hundreds of compounds with known activities against a range of therapeutically relevant targets and were extracted from the MDDR.[47,48] These datasets will also be used in this study.

1.3.2 Impact of Dataset Composition on Validation Results

Although the validation and calibration of VS methods on benchmark datasets is now an established methodology, a number of problems arise from its empirical nature. In a study on the evaluation of the docking program GOLD, Verdonk et al. have shown that results are highly influenced by the composition of the dataset of decoys.[72] They proved that if the background significantly differs from the set of actives regarding "low dimensional" properties like molecular weight or number of hydrogen bond donors/acceptors, it may lead to "artificial enrichment". That is the classification is actually caused by the differences in bulk or global molecular properties rather than specific interactions with the target. They conclude that focusing the library of inactives to the same range of low dimensional properties as the actives is essential for the results of VS validation to be representative. Recently DUD ("Directory of useful decoys"), a collection of validation sets for molecular docking seeking to fulfill these requirements, has become available for public use.[67]

As mentioned above, Bender and Glen showed that the random ranking hypothesis underlying the calculation of enrichment factors often leads to overoptimistic estimations of a method's performance.[58] They suggest normalizing the RTR of any VS method by the RTR of a so-called "dumb" descriptor like molecular weight or atom counts, which is basically a vectorized form of the chemical sum formula. Since the sum formula is highly correlated with molecular weight and hydrogen bond donor/acceptor counts, this approach implicitly covers many of the phenomena discussed by Verdonk et al. and can effectively be used as a negative control when validating VS methods utilizing similarity searching.

While the works by Verdonk et al. and Bender et al. precisely highlight the effect the composition of the background dataset has on the outcome of VS validation, three papers by Good and coworkers concluded that datasets of actives extracted from databases constructed from drug discovery projects, such as the MDDR, are prone to over-representation of certain scaffolds or chemical entities.[73–75] It was shown, that unless the benchmarking datasets and the background are chosen with care, the figures of

merit for ligand based virtual screening may be over-optimistic due to the so-called "analogue bias".

Recently, Vogt and Bajorath proposed a measure of divergence between the descriptor distributions of the benchmark dataset and the background that relates to VS performance and can thus be used to estimate performance rates in descriptor-based virtual screening.[76,77] Their approach is based on a number of assumptions about the descriptor distribution. As opposed to this, the method presented in the course of this study is non-parametric, i.e. can deal with arbitrary distributions of benchmark dataset and background.

1.3.3 Chemical Space

Many of the concepts discussed so far already implied an interpretation of chemical data in the context of analytical geometry and spatial relations. The application of distances[29] and other geometric analogies to chemical datasets and databases has proven to possess enormous descriptive power. Therefore, the discipline of chemoinformatics has widely adopted the concept of so-called "chemical space". Basically, chemical space is defined as the multitude of all molecules that can theoretically exist given the rules of chemistry. Current estimations suggest that this comprises about 10^{23} theoretically feasible molecules.[78] Inside this huge chemical universe, several subspaces exist that are defined by a certain property common to all molecules in this subspace.[79] Well known examples include "drug-like" chemical space or "peptide space".

A very important class of chemical subspaces are those formed by all compounds active against a certain molecular target. These so-called activity spaces are central to the analysis and design of VS benchmark datasets. A good benchmark dataset should cover the activity space of its target comprehensively.

In the context of descriptor based ligand based virtual screening, chemical space can be defined more stringently. For any class of molecular descriptor, chemical space is defined as the set of coordinates that can be occupied by vectors belonging to chemical compounds represented by this descriptor. Analogously, the activity space of target X is

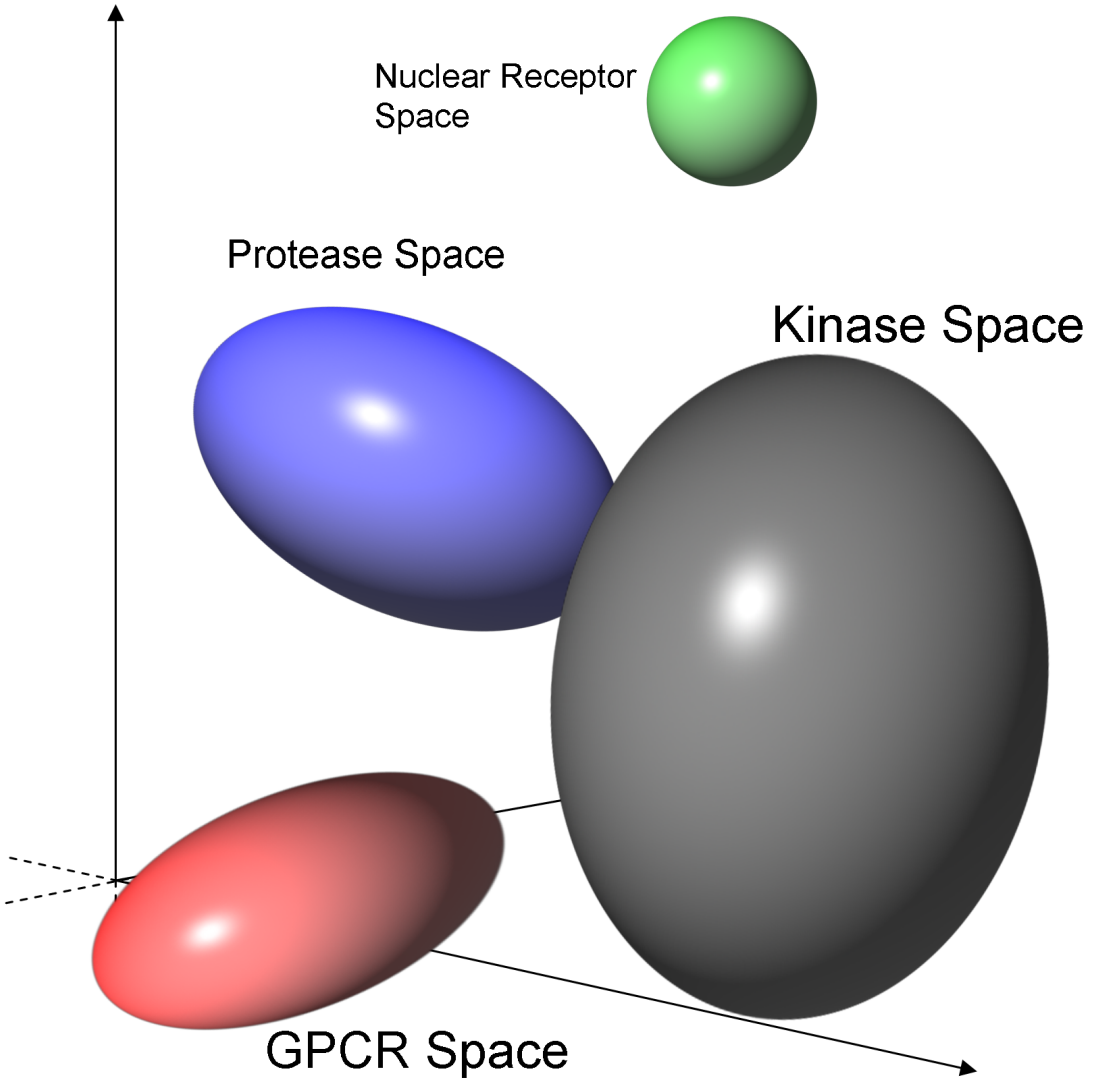


Figure 1.10: Cartoon representation of the continuum of chemical space and the subspaces occupied by molecules with specific biological activities. One possible way of defining subspaces is the grouping of molecules by the gene families of their targets, e.g. proteases (green), GPCRs (red), kinases (grey) or nuclear receptors (green). The activity space of a particular GPCR would itself be a subspace of GPCR space. *Figure adapted from Lipinski et al.*[79]

the subset of coordinates occupied by compounds with activity against X. The concept of chemical space and the spatial description of the extent and shape of various subspaces will be central to this study.

1.3.4 Benchmark dataset topology

When comparing the benchmark datasets proposed by Hert et al.[47,48] with datasets of actives that form the input of real-life virtual screening campaigns, the most striking feature is their size. Whereas the benchmark datasets consist of several hundreds to more than thousand substances, usually only a small number ($\sim 10-20$) of active substances are available at the beginning of a VS campaign, rendering the benchmark datasets redundant with respect to real-life VS. Furthermore, the potential presence of large scaffold families in the datasets introduces analogue bias, as stated above. These phenomena and their variations across datasets manifest themselves as spread, self-similarity, patchiness and clustering of the datasets' representation in descriptor space. We will refer to this mapping of dataset composition into descriptor space as the dataset topology. Both, the calculation of molecular similarity and the subsequent ranking of the screened database, which are the most crucial steps in the conduction of virtual screening, are based on the respective molecules' representation by descriptors. It is therefore reasonable to quantify analogue bias, redundancy and other phenomena influencing VS performance by their representation in descriptor space, i.e. by dataset topology.

Chapter 2

Impact of Dataset Topology on VS Validation

2.1 Objectives

The works of Verdonk et al.,[72] Bender et al.[58] and Good et al.[73–75] described the impact of benchmark dataset composition on VS validation results in a qualitative manner. Important termini like "artificial enrichment" and "analogue bias" were introduced, but lack means for their quantification. As a consequence, no quantitative correlation between the occurrence and degree of artificial enrichment or analogue bias in a dataset and the results of VS validation has yet been demonstrated. Therefore, the first objective of this study is to develop a methodology for the quantification and description of dataset composition. The methodology will be based on a spatial analysis of the datasets' topology in descriptor space. The extensive impact of benchmark dataset composition on VS validation results will be demonstrated by application of the methodology on literature benchmark datasets combined with retrospective VS simulations.

2.2 Methods

2.2.1 Methodological Strategy

The basic idea of this part of the study is to sample sub-sets with different topologies from the well established Hert-Willett benchmark datasets.[47,48] (see Section 1.3.1) By comparing the results of retrospective VS simulations on these sub-samples, the influence of dataset topology on the validation results can be observed. Various sampling strategies are employed to generate archetypal sub-samples from the literature benchmark datasets: (i) maximum diversity subsets, (ii) space filling samples and (iii) subsets with minimum intra-set diversity. The analysis of the varying VS performance on these prototype datasets allows us to assess if and to what extent dataset topology affects the validation of virtual screening. Here, it is essential to ensure that no other factors influence the outcome of the VS runs, which is achieved by a careful design of the experimental setup. The magnitude of factors of variance that can not be controlled by the experimental setup is estimated by bootstrapping methods. (Figure 2.1) Spatial statistics techniques are introduced here to explore and quantify the effect of benchmark dataset topology in more detail. In order to get an idea how encoding by different descriptors influences dataset topology in descriptor space, two kinds of descriptors are used for the experiments: MOE molecular descriptors and "simple" descriptors acting as a negative control. After using artificial sub-samples to show that dataset topology has an effect on the outcome of VS validation and that it can be quantified by spatial statistics methods, the methodology is applied to the complete benchmark datasets.

2.2.2 Datasets

The datasets as described by Hert et al.[47,48] and two additional datasets containing inhibitors of angiotensin converting enzyme (ACE) and acetylcholineesterase (AChE) were extracted from the MDDR[69] (MDL Drug Data Report) by their activity indices. An overview of the datasets is provided by Table 2.1. The remaining 93925 molecules in the MDDR that did not belong to at least one of the activity classes, were assumed to be inac-

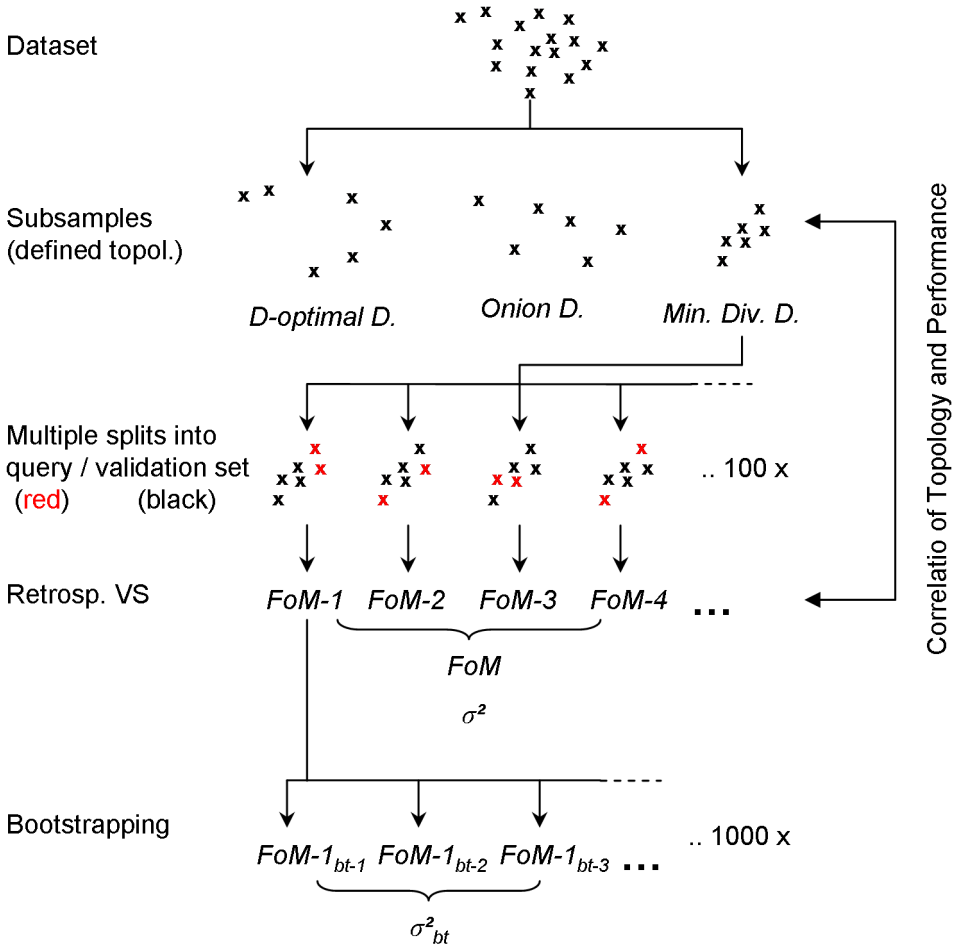


Figure 2.1: Schematic outline of the experimental setup. Sub-samples with controlled topology are extracted from literature benchmark datasets. 100 random splits of the sub-sample into query and validation set are generated for retrospective VS. The arithmetic mean and the variance of the figure of merit (FoM) under scrutiny over all splits estimates the VS performance on the sub-sample and its statistical error, respectively. Bootstrapping is used to determine to what extent this statistical error is caused by factors other than topology.

Table 2.1: Hert-Willett Benchmark Datasets with Activity Against the Specified Targets Extracted from the MDDR.

Activity	Dataset size
Angiotensin converting enzyme (ACE) inhibitors	355
Acetylcholine esterase (AChE) inhibitors	701
Angiotensin II Type 1 Receptor Blockers	943
Cyclooxygenase (COX) inhibitors	636
D2 antagonists	395
HIV protease inhibitors	750
5HT1A agonists	827
5HT3 antagonists	752
5HT reuptake inhibitors	359
Protein kinase C (PKC) inhibitors	452
Renin inhibitors	1130
Substance P inhibitors	1246
Thrombin inhibitors	803

tive and were used as decoys as described by Hert et al.[47,48] This excludes potential overlap between the sets of actives as a factor distorting validation results. SD-Files[80] of all datasets and the background were cleaned from small fragments like counter ions and solvents using MOE (Molecular Operating Environment)[34] and converted to 3-dimensional structures with CORINA.[81]

2.2.3 Descriptors

For all datasets and the set of decoys, MOE molecular properties descriptors[34] were computed. Properties are grouped into three classes: 2D (computed from the 2-dimensional topology of a molecule), i3D (properties that depend on internal 3-dimensional coordinates) and x3D (calculated from a grid surrounding the 3-dimensional structure of the molecule). Since the x3D class depends on a common spatial frame of reference for all molecules, i.e. an alignment of all molecules in a spatial grid, which is not feasible for VS applications, only the 2D and i3D classes were used for experiments. The numerical values of properties in MOE descriptors have significantly different ranges. Molecular weight, for instance, typically varies between 0 and $\sim$ 1000, whereas logP often has a value roughly between -1 and 5 for drug-like molecules. Therefore the data matrix con-

sisting of MOE descriptor vectors for the complete MDDR was autoscaled columnwise by subtracting the mean and dividing by the standard deviation of each column. Columns whose properties had constant values for the complete MDDR were removed from the matrix. After this pre-treatment, the matrix had a dimensionality of 180. In order to reduce noise in the descriptor matrix, principal components analysis (PCA)[82] was applied. An analysis of the resulting eigenvalues showed that $> 99\%$ of the total variance could be explained by the first 54 components. Thus the first 54 scores from the PCA were used as the final descriptors for the VS simulations.

Following the approach suggested by Bender and Glen,[58] the database was also encoded by negative control "simple" descriptors, which consisted of the respective counts of all atoms, heavy atoms, boron, bromine, carbon, chlorine, fluorine, iodine, nitrogen, oxygen, phosphorous and sulfur atoms in each molecule as well as the number of H-bond acceptors, H-bond donors, the logP, the number of chiral centers and the number of ring systems. These properties were generated from the SD-Files of the datasets using OpenEye BABEL3[83] (atom counts) and OpenEye FILTER[84] (acceptors, donors, logP, chiral centers, ring systems). The respective output files were parsed with an in-house PERL script to provide the final descriptors.

These descriptors are central to this study, because they capture all molecular properties associated with analogue bias and artificial enrichment.[72–75] They are a valuable tool for the detection and prevention of bias in VS validations introduced by benchmark dataset composition. Thus, they will be utilized extensively throughout the course of this study. For the sake of brevity, they will be denoted simple descriptors from now on in the text.

The simple descriptors for the MDDR datasets used in this Chapter were not autoscaled, since it was deemed important to preserve their chemical meaning. This facilitates a comparison of MOE descriptors with crude molecular properties. Autoscaling of descriptors is most important for the design of datasets with specific topological properties, such as in Sections 2.2.4 and 3.2.10. For such purposes, it is desirable that the design is determined equally by all dimensions of a descriptor vector. In this Chapter,

all design steps are based on PCA scores of MOE descriptors, which are centered, autoscaled and orthogonal. Therefore an autoscaling of simple descriptors was not essential. In Chapter 3, Section 3.2.10, however, simple descriptors are used as the basis for dataset design. Accordingly, the simple descriptors used in Chapter 3 were centered and autoscaled. (see Section 3.2.8)

2.2.4 Sub-Samples with Defined Topology

Various methods exist to generate subsets of compounds based on a descriptor representation of the original dataset. Based on the MOE representation of the data, from each of the benchmark datasets a subset of $k = [50, 100, 150, 200, 250, 300]$ compounds was generated for each of the three following sampling strategies. *D-optimal design*[85,86] was used to provide subsets with the maximum intra-set diversity for the respective number of compounds. For the generation of subsets sampling the respective dataset in a space filling manner, *D-optimal onion design*[87] was applied. Around the center of mass of each dataset 5 shells were defined, of which each contained 20% of the data. In contrary to shells of equal distance, this approach ensures the presence of an adequate number of datapoints in each shell in spaces of high dimensionality.[88,89] In order to reflect the datasets' density distribution in the sub-samples, an equal number of $k/5$ compounds were chosen from each shell by the D-optimality criterion. Both, D-optimal design and D-optimal onion design were implemented using the Statistics Toolbox of The Mathworks MATLAB 7[90]. Subsets with a minimum sum of intra-set all against all distances, i.e. by a *Minimum Diversity Design* were generated using an in-house row exchange algorithm[91] also implemented in MATLAB 7. All three sampling strategies are deterministic, i.e. for a given k one optimum sub-sample from the original dataset is selected. By this procedure three prototypes of sub-samples were created for every k: (i) A "worst-case scenario" for which VS using MOE descriptors would be very difficult was generated by the maximum diversity criterion using D-optimal design. (ii) An intermediate case with active compounds equally spread across MOE descriptor space with varying density depending on k was generated by the onion design approach. (iii) Finally, a "best-case scenario"

with multiple active compounds concentrated in a tight cluster resulted from the minimum diversity design. In order to observe how dataset topology and VS performance are affected by different descriptor representations, the identical sub-samples, i.e. the same compounds chosen by the different design strategies for MOE descriptors, were compiled for the simple descriptor representation. This excluded sub-sample composition as a factor of variance when comparing the performance of different descriptor representations.

2.2.5 Retrospective Virtual Screening Simulations

A variety of methods is available for the conduction of ligand based virtual screening. It was not the goal of this study to figure out the descriptor, similarity measure, or searching method that presumably generates best results for virtual screening, but to determine the influence of dataset topology on method performance. Therefore we settled for one type of searching procedure and one similarity measure, which we kept constant for all experiments. For similarity searching with multiple query molecules, data fusion according to the "MAX"-rule[47,48] has proven to be very powerful (see Section 1.1.3.2) and was therefore used for all experiments presented in this Chapter. All simulations of virtual screening described in this Chapter were carried out using ten active compounds chosen randomly from the sub-samples described above as query and pooling the remaining actives with the decoys. Similarity was measured by the Euclidean distance and the respective validation sets were ranked accordingly. This was repeated 100 times to assess the variability of the results and to obtain a mean value that is not affected by the random choice of the query molecules. To obtain an estimate of the statistical error of the respective rankings, 1000 bootstrapping runs were carried out on the ranking resulting from each of these query / validations set combinations, randomly leaving out 20% of both the actives and the background compounds in each run, respectively.

2.2.6 Figures of Merit (FoM) for Virtual Screening Performance

The ability for early recognition of active compounds was measured by the mean fraction of retrieved actives in the first percent of the ranked validation set $RTR_{1\%}$. Additionally,

the area under the receiver operating characteristic curve (AUC_{ROC}) was determined for all rankings in order to rule out any bias introduced by the hard 1% cutoff of the $RTR_{1\%}$. The mean Retrieval Rates and mean areas under the receiver operating characteristic curves obtained from the 100 random query / validation set splits, which were generated for each dataset sub-sample, will be denoted $mean(RTR_{1\%})$ and $mean(AUC_{ROC})$ throughout the text.

2.2.7 Variance Decomposition for Figure of Merit Statistical Errors

The standard deviation of both $RTR_{1\%}$ and AUC_{ROC} is often used to estimate the statistical error of VS validation results. In the context of this study, it is desirable to isolate the component of error that is caused by the topology of the dataset used in the respective validation experiment. In their recent paper, Truchon and Bayly not only introduced a new metric for the validation of VS methods, but also provided a remarkably concise and comprehensive analysis of the factors influencing the variance of VS rankings and the resulting figures of merit.[60] According to them, the variance is mainly influenced by the following parameters: N, the absolute number of decoy molecules and R_a, the fraction of actives in the ranking (i.e. size of the dataset of actives). Furthermore, the "goodness" of a ranking itself (denoted λ by Truchon and Baily) has considerable impact on the variance of the figures of merit. This is caused by the fact, that for a very "successful" screening run the actives are far less spread over the ranking, since they are concentrated at its beginning. Furthermore, they discuss a "saturation effect" that affects metrics measuring early recognition if the number of active compounds is higher than the number of ranks in the part of the ranking that is considered "early".

In our setting, the number of decoys was kept constant ($N = 93925$) for all experiments and can therefore be ruled out as a factor affecting the results of the VS simulations. Since the size of the biggest subsets of actives used here is 300, which is clearly smaller than the number of ranks in the first percent of the rankings ($\backsim$ 940), the RTRs reported here are not subject to the saturation effect. ROC metrics have no early recognition features and consequently are not prone to the respective saturation effects. There are, however, two

parameters that can not be kept constant in this experimental setting: λ: This parameter is used by Truchon and Bayly to denote rankings with varying VS performance. However, the rankings discussed in their paper are derived from sampling a model probability density function rather than real VS runs. In our setting, the "goodness" λ of the ranking can not be determined before the experiment. R_a: The sampling strategy for the design of subsets with different topology requires subsets of actives with differing sizes. The magnitude of the variance component introduced by changing R_a and λ can be estimated by the bootstrapping procedure described above. As indicated in Figure 2.1, 1000 bootstrap samples were drawn from each ranking obtained from a particular query / validation set split. The variance $\sigma^2_{bt,i}$ with $i = 1..100$, is an estimate of the statistical error caused by the particular combination of λ and R_a for each of the 100 splits. According to the law of total variance for experiments with a nested design the overall variance σ^2, can be decomposed to yield a variance component σ^2_{top} associated with dataset topology:[92]

$$\sigma^2 = \sigma^2_{top} + mean(\sigma^2_{bt,i}); \tag{2.1}$$

which can be written as:

$$\sigma^2_{top} = \sigma^2 - mean(\sigma^2_{bt,i}); \tag{2.2}$$

With all other factors constant, the only factor affecting σ^2_{top} is the dataset topology. A corrected standard deviation providing an estimate of the statistical error of the validation results associated with dataset topology can then be calculated as:

$$std_{top}(FoM) = \sqrt{\sigma^2_{top}};\ FoM = RTR_{1\%}\ or\ AUC_{ROC} \tag{2.3}$$

2.2.8 Spatial Statistics Analysis of Chemical Datasets

2.2.8.1 Basic Categories of Dataset Topology

One major goal of this study is to describe the topology of chemical datasets or subsamples of datasets by the relative positions of the datapoints in chemical space to which

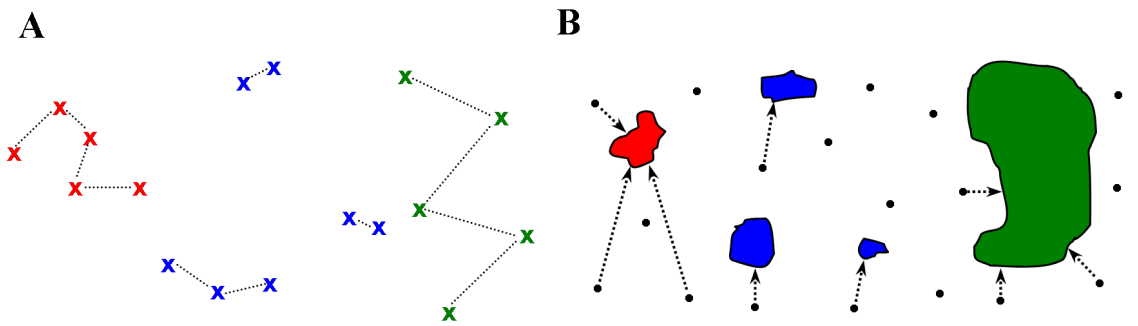


Figure 2.2: Representative datasets and their characterization by Refined Nearest Neighbor Analysis. (A) Nearest neighbor distances in concentrated (red) and patchy (blue) sets are smaller than in dispersed sets (green). However, patchy and concentrated sets can not be distinguished by nearest neighbor distances, since these are not affected by the presence of multiple clusters. (B) By "flooding" the analyzed space with random points and measuring their distances to the nearest event, gaps in the data can be detected and the overall spread of the sets can be quantified, thereby differentiating patchy from concentrated sets.

the compounds in the dataset are projected by the application of a structural descriptor. In the terminology of spatial statistics, the position of a compound belonging to the sample under examination is an "event". On the other hand, "points" denote arbitrary coordinates in chemical space. Regarding virtual screening, there are three basic categories of topology for a set of events. (Figure 2.2) (i) "Concentrated" sets consist of a single dense cluster, well separated from the rest of chemical space. (ii) "Patchy" sets are composed of multiple dense but separated clusters. (iii) "Dispersed" sets are regularly distributed in chemical space, with comparatively large event-event distances. It should, however, be kept in mind that this rather coarse categorization of dataset topologies is mainly used here for the illustration of the basic properties of the spatial statistics functions presented below. Real chemical datasets will usually incorporate all kinds of nuances and combinations of the three basic categories of dataset topology. In the following two sections a methodology within the framework of spatial statistics will be developed, that facilitates the detection and identification of all major types of topologies and the quantification of "dataset clumping" in chemical datasets. Refined Nearest Neighbor Analysis[93,94] is a set of spatial statistics methods that is especially useful for the analysis of large datasets. Its mathematical foundations will be introduced in Section 2.2.8.2. Different approaches for the implementation of the underlying functions will be discussed in Section 2.2.8.3.

2.2.8.2 Refined Nearest Neighbor Analysis: Mathematical Foundations

Refined Nearest Neighbor Analysis is a mathematical framework for the analysis of mapped point patterns. It is based on the calculation of two functions from the position of points and events (Figure 2.2): $G(t)$ is the proportion of events for which the distance to the nearest neighbor is less than t. $G(t)$ is called the "nearest neighbor function" and is a cumulative probability distribution of the distance of any event to its nearest neighbor event. Using $G(t)$, it is possible to distinguish dispersed from concentrated and patchy sets (green vs. blue and red in Figure 2.2A). In some cases, it is however not possible to differentiate between patchy and concentrated sets (blue vs. red in Figure 2.2A). Since only the nearest neighbor of each event is considered, the spacing between several dense clusters has no effect on $G(t)$, because the nearest neighbor of any event will always be located in the same cluster. As a consequence, $G(t)$ is neither sensitive to the presence nor to the spacing of multiple clusters. This distinguishes the nearest neighbor function from approaches based on average intra-set distances or methods based on more than one neighbor. Let n be the number of events, then $G(t)$ is given as:

$$G(t) = \frac{\sum_{n} I_t(\mathbf{i},\mathbf{j})}{n}; \tag{2.4}$$

with $I_t(\mathbf{i},\mathbf{j}) = 1$ if the distance of event $\mathbf{i}$ to its nearest neighbor $\mathbf{j}$ is smaller than t. Representative graphs of $G(t)$ for concentrated, patchy and dispersed sets are shown in Figure 2.3A.

In order to distinguish between concentrated and patchy sets, a large number of points are sampled randomly from chemical space. $F(t)$ is the proportion of these points for which the distance to the nearest event is less than t. (Figures 2.2B, 2.3B) $F(t)$ is a cumulative probability distribution of the distance from a randomly chosen point to the nearest event and is often called the "empty space function", because it is sensitive to gaps in the data. For a patchy set, the average distance from a random point to the nearest event will be smaller than for a concentrated set. Depending on the number and the degree of separation between the clusters less chemical space is unoccupied for patchy sets. On

the other hand, because $F(t)$ does not take into account event-event distances, it can not differentiate between dispersed and patchy sets, if the clusters in the latter are far apart. Figure 2.3B provides representative graphs of $F(t)$ for concentrated, patchy and dispersed sets. Let m be the number of random points, then $F(t)$ is given as:

$$F(t) = \frac{\sum_{m} I_t(\mathbf{j},\mathbf{i})}{m}; \tag{2.5}$$

with $I_t(\mathbf{j},\mathbf{i}) = 1$ if the distance of point $\mathbf{j}$ to its nearest neighbor event $\mathbf{i}$ is smaller than t.

Incorporating the information of both functions into one equation, it is possible to differentiate and quantify all types of dataset topology:

$$S(t) = F(t) - G(t); \tag{2.6}$$

Figure 2.3C shows typical graphs of $S(t)$ for the major topology types. By summing up the values of $S(t)$ over a range of distances t_i, i.e. by numerical integration, the clustering behavior of the set can be estimated by a single scalar:

$$\Sigma S = \sum_i S(t_i); \tag{2.7}$$

where t_i represents a series of distances. The scalar value of ΣS provides a quick and easily interpretable estimate of a set's clumping or dispersion, with values of $\Sigma S < 0$ indicating clumping and values of $\Sigma S > 0$ indicating dispersion. Most real-life chemical sets can not be strictly assigned to one of the basic categories of topology, but differ in the number of clusters, their respective density and scaling. This is perfectly reflected by the scalar ΣS, which provides a quantitative measure for the degree of clumping in a sample. (Figure 2.3D)

Under certain conditions, it can be useful to calculate scalars analogous to ΣS from $G(t)$ and $F(t)$. Summing up the values of both functions over a range of distances t_i yields two scalars

$$\Sigma G = \sum_i G(t_i); \tag{2.8}$$

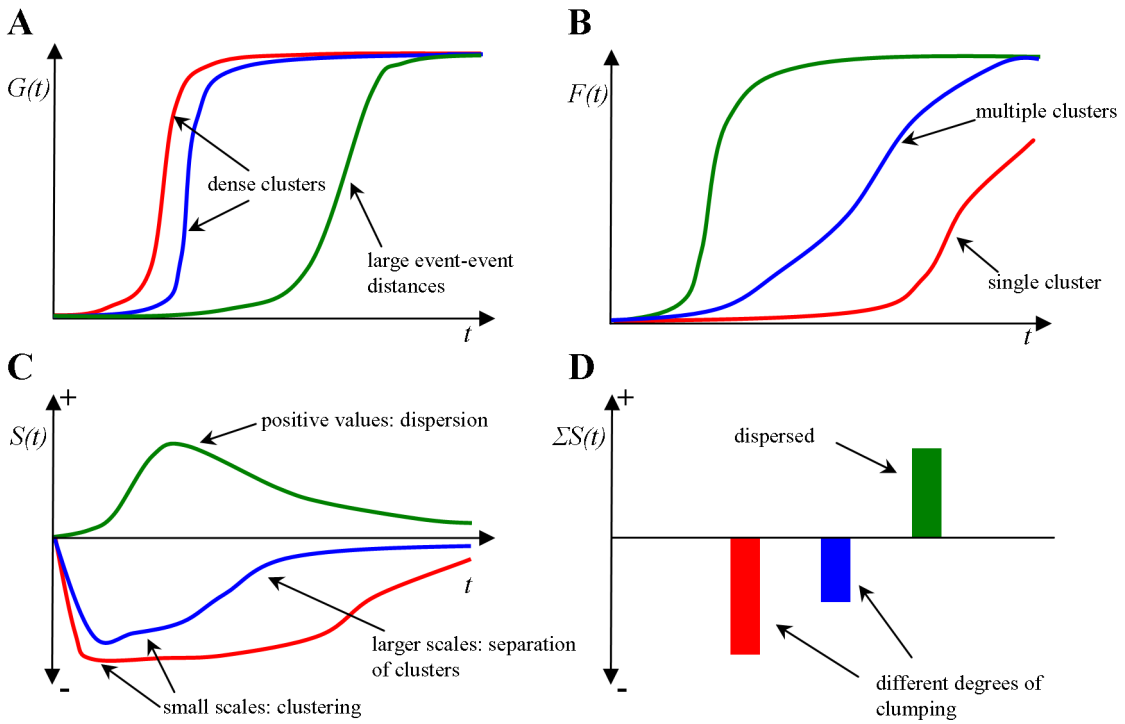


Figure 2.3: Exemplary graphs of Refined Nearest Neighbor Analysis functions. (A) The graphs of $G(t)$ for both concentrated (red) and patchy (blue) sets feature a steep ascent at small values of t, caused by clusters of high density. Dispersed sets (green) can easily be differentiated. (B) Patchy sets (blue) can be distinguished from concentrated ones (red) by the earlier rise in $F(t)$. For large separation between the clusters, the graph for patchy sets would converge to the graph for dispersed ones (green). (C) Using $S(t)$ the topology of sets can be characterized unambiguously. Whereas dispersed sets (green) are marked by positive values for $S(t)$, different types of clustering exhibit characteristic curves in the negative region. (D) ΣS, the area under the curve of $S(t)$, provides an easily interpretable estimate of the degree of clumping or dispersion of a set.

$$\Sigma F = \sum_i F(t_i); \tag{2.9}$$

that are robust estimates for the self-similarity (ΣG) in the sample of events, i.e. analogue bias, and the separation between points and events (ΣF), i.e. artificial enrichment. Large values of ΣG indicate a high level of self-similarity among events, whereas small values of ΣF indicate a high degree of separation between points and decoys.

2.2.8.3 Refined Nearest Neighbor Analysis: Implementation

The implementation of both $G(t)$ and $F(t)$ requires a prior definition of the range of distances t_i at which they are calculated. This range of values t_i depends on the particular

application of Refined Nearest Neighbor Analysis and must be determined empirically. For the analyses based on the Hert-Willett datasets presented in this section, preliminary experiments identified $t_i = [0.01, 0.02...12]$ as the best choice.

Later experiments (see Section 3.2.9.1) showed that the upper boundary of t_i should be set to a value $t_{max} = c * d_{mnn}$ and that t_i should be incremented in fractions of d_{mnn}. Here, d_{mnn} is the median of the distance to the nearest neighbor for all molecules in the respective descriptor space (the complete MDDR encoded by MOE and simple descriptors respectively). c is a constant depending on the datasets under examination. The median nearest neighbor distance in the complete MDDR is $d_{mnn} \approx 5$ for both, MOE descriptors and non-scaled simple descriptors, rendering the choice of

$$t_i = [0.01, 0.02...12] = [0, \frac{1}{1200}, ..., 1] * 2.4 * d_{mnn}$$

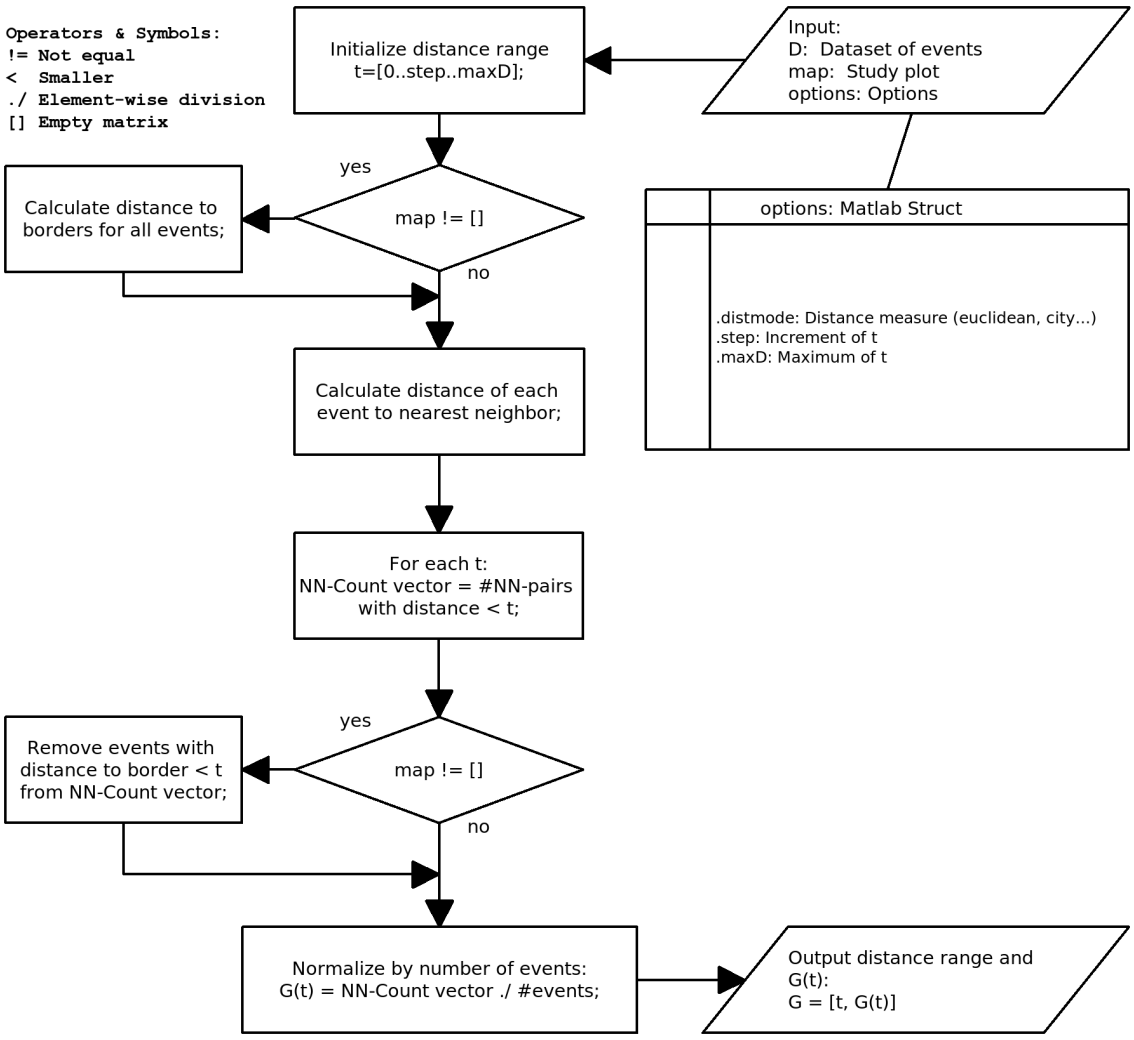
consistent with these later findings.

With t_i defined, implementing $G(t)$ was straightforward. For any given dataset or dataset sub-sample, the distance of each event (active compound) to its nearest neighbor event was determined and $G(t)$ was calculated according to Equation 2.4 using Algorithm 2.1. (For the source code see Appendix D.2.1)

Refined Nearest Neighbor Analysis was originally developed for the analysis of spatial patterns in ecology and forestry.[93,94] In these disciplines, the examined spatial phenomena can usually be represented by points on 2-dimensional, finite maps, so-called "*study plots*". On such plots, the dataset of random points can be generated by Monte Carlo samplings from a uniform 2-dimensional distribution of coordinates.

In the context of chemical space however, the implementation of $F(t)$ (Algorithm 2.2) faces the problem, that the descriptor spaces used in virtual screening - and in this study - are of rather high dimensionality and non-finite. High dimensional spaces are subject to the "empty space phenomenon", i.e. they are inherently sparsely populated.[89] As a consequence, if $F(t)$ was calculated based on random points uniformly distributed in such a high dimensional space, its value would be dominated by the empty space inherent to the dimensionality, not by the gaps in the data. Thus, any sampling of points must ensure

Algorithm 2.1 Calculation of $G(t)$ For a Dataset D. Also Applicable for Finite Study Plots (Maps).



that these points lie in the portion of chemical space actually populated by compounds, but nevertheless provide representative coverage.

This was achieved using three approaches for the generation of sets of random points (see Algorithm 2.2, Source code: Appendix D.2.2): (i) Bootstrapping from the complete MDDR: 10000 compounds were randomly chosen from the union of the background and all benchmark datasets. (ii) Boostrapping from the set of decoys: 10000 compounds were chosen by random from the set of inactives. (iii) Convex pseudo-data: Following the approach described by Breiman et al.[95] 10000 pseudo-data points were generated from 20000 compounds chosen randomly from the MDDR. Briefly, from two datapoints $\mathbf{x_1}$, $\mathbf{x_2}$ a new pseudo-datapoint $\mathbf{x_3}$ is generated by selecting a random number v from the interval $[0,1]$. Then $\mathbf{x_3}$ is given by the linear combination:

$$\mathbf{x_3} = v * \mathbf{x_1} + (1-v) * \mathbf{x_2} \tag{2.10}$$

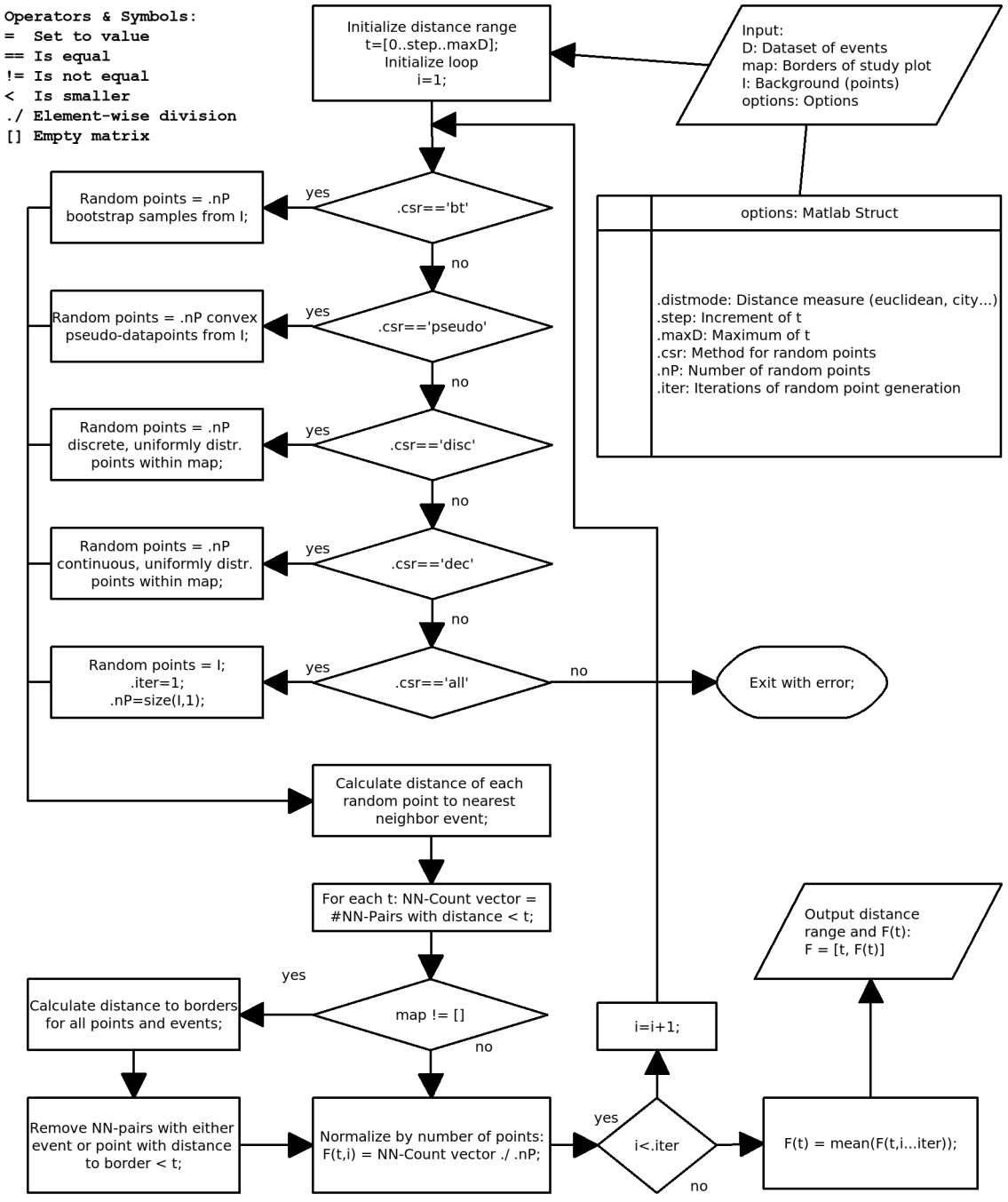
Thereby, artificial pseudo-datapoints are created that occupy the same region of chemical space as the original population of datapoints.

For each sub-sample under scrutiny, this was repeated 20 times and $F(t)$ was calculated using Equation 2.5 with $t_i = [0.01, 0.02...12]$ for all samples of random points. The final value of $F(t)$ was determined as the arithmetic mean for all 20 sets of random points for each sampling method. Results for $F(t)$ were equal within the margin of statistical error for all methods of random point sampling. Therefore, all results will be reported for $F(t)$ with bootstrapping from the complete MDDR. An additional advantage of this procedure is, that there is no need for terms of edge correction in $F(t)$ and $G(t)$, which are necessary in traditional spatial statistics applications on finite maps. $S(t)$, ΣS, ΣG and ΣF resulted from $G(t)$ and $F(t)$ as given by Equations 2.6, 2.7, 2.8 and 2.9.

2.2.9 Visualization of Topology by Self-Organizing Maps (SOMs)

Self-Organizing Maps (SOMs)[96] are a special type of artificial neural networks, that project high-dimensional data onto a two-dimensional lattice while preserving the topology of the input data space. It is obvious, that this feature makes SOMs the ideal tool for

Algorithm 2.2 Calculation of $F(t)$ for a Dataset D and a Background Dataset I.



the visual perception of dataset topology. Comprehensive reviews of the SOM algorithm and its variations exist in the literature.[97–101] All SOMs used here were generated and analyzed using the SOM-Toolbox 2.0[102] in The Mathworks MATLAB 7.[90] Map topology was chosen to be non-toroidal with a rectangular lattice. Since SOMs were only used for visualization supporting a more detailed analysis of dataset topology by Refined Nearest Neighbor Analysis, a map grid of 20 x 20 units was considered sufficient. For training, the batch algorithm as implemented in the SOM-Toolbox was used.

2.2.10 Measures of Correlation

One of the goals of this study is to quantify the influence of dataset topology on the outcome of VS validation experiments by the degree of correlation between scalar measures of dataset topology and VS performance. (see Section 2.1) These measures span numerically different ranges, namely $[-\infty, \infty]$ for measures of topology and $[0, 1]$ for measures of VS performance. Moreover, none of the examined quantities can be safely assumed to be normally distributed, nor is there any hard evidence for a linear relationship of VS performance and dataset topology. Therefore, the Spearman rank correlation coefficient ρ[103] was used as the measure of correlation throughout this study. In contrast to the widely used Pearson-product moment correlation coefficient,[92] the Spearman rank correlation coefficient can be utilized to detect a correlation between two variables X and Y even if their relationship is non-linear.

For the calculation of the Spearman correlation coefficient, two sets of data X, Y are first converted to ranks x, y by ordering the datapoints:

$$x_i = rank(X_i); \tag{2.11}$$

$$y_i = rank(Y_i); \tag{2.12}$$

Where X_i, Y_i constitute the i^{th}datapoints in the two datasets and x_i, y_i constitute the corresponding ranks. Using these ranks x_i, y_i, the calculation of the Spearman correlation

coefficient ρ is equivalent to the calculation of the Pearson correlation coefficient on the rank tranfsormed data:

$$\rho = \frac{n\left(\sum_i x_i y_i\right) - \sum_i x_i \sum_i y_i}{\sqrt{n\left(\sum_i x_i^2\right) - \left(\sum_i x_i\right)^2}\sqrt{n\left(\sum_i y_i^2\right) - \left(\sum_i y_i\right)^2}} \tag{2.13}$$

with n the number of samples in the sets X_i and Y_i, respectively. ρ takes values in the interval $[-1,1]$, with the sign of ρ indicating the direction of the correlation. Here, a value of 1 indicates a perfect positive correlation among the two sets of ranks x_i and y_i and a value of -1 indicates a perfect negative, or *anti*-correlation, whereas a value of 0 is an indicator of no correlation at all.

The value of ρ, at which the null-hypothesis of no correlation can be rejected safely, i.e. with a confidence level of 95%, can be calculated using random permutation tests. For all measures of topology employed in this study only one respective direction of correlation with VS performance, i.e. positive or negative correlation, is reasonable. Therefore all levels of significance for ρ given in this study were calculated based on the *one-sided* 95% confidence interval of ρ for rejecting the null hypothesis of no correlation. Depending of the expected algebraic sign of the correlation, the boundary of this interval was calculated as the 95^{th} percentile (right tail) for positive correlations and as the 5^{th} percentile (left tail) for negative correlations of a distribution of correlation coefficients ρ_0 generated by 100000-fold random permutation of the respective samples.[104] Table 2.2 provides an overview of the signs of the respective correlation coefficients with VS performance and the respective percentiles utilized as boundaries of the confidence intervals.

2.3 Results

2.3.1 Characterization of Benchmark Dataset Sub-Samples by Refined Nearest-Neighbor Analysis

Using the sampling procedure described above (see Section 2.2.4), six sub-samples of different size were generated for every sampling strategy from any of the benchmark datasets. Doing so, a total of 234 (6 sample sizes $\times$ 3 sampling strategies $\times$ 13 datasets) sub-samples were generated featuring all types and nuances of topology, ranging from dispersed (D-optimal Design) and patchy sub-samples (Onion Design) to concentrated ones (Minimum Diversity Design). For all samples, topology was characterized using Refined Nearest Neighbor Analysis. The differences in sub-sample topology were well reflected by $G(t)$, $F(t)$, $S(t)$ and ΣS. The degree of clumping for each sub-sample as quantified by ΣS in MOE and simple descriptor space is given by Tables A.1 and A.2 (Appendix). An example for the visualization of sub-sample topologies and the results of Refined Nearest Neighbor Analysis for four sub-samples from the dataset of Thrombin inhibitors is provided by Figures 2.4 and 2.5, respectively.

2.3.2 Correlation of VS Performance and Dataset Clumping

As described in Section 2.2.5, retrospective VS validation experiments were carried out for 100 query / validation set splits of each sub-sample using MOE-PCA and simple

Table 2.2: Algebraic Signs of Correlation Coefficients of Different Measures of Dataset Topology with VS Performance.

Measure of Topology	Sign of Correlation Coefficient	Boundary of Confidence Interval
ΣS	-	5^{th} percentile[a]
ΣG	+	95^{th} percentile[a]
ΣF	-	5^{th} percentile[a]
${g_{med}}^b$	+	95^{th} percentile[a]
${f_{med}}^b$	-	5^{th} percentile[a]
AvD^b	-	95^{th} percentile[a]

[a)] of a distribution of 100.000 correlation coefficients ρ_0 generated by random permutations of the examined data.

[b)] (see Section 2.3.4)

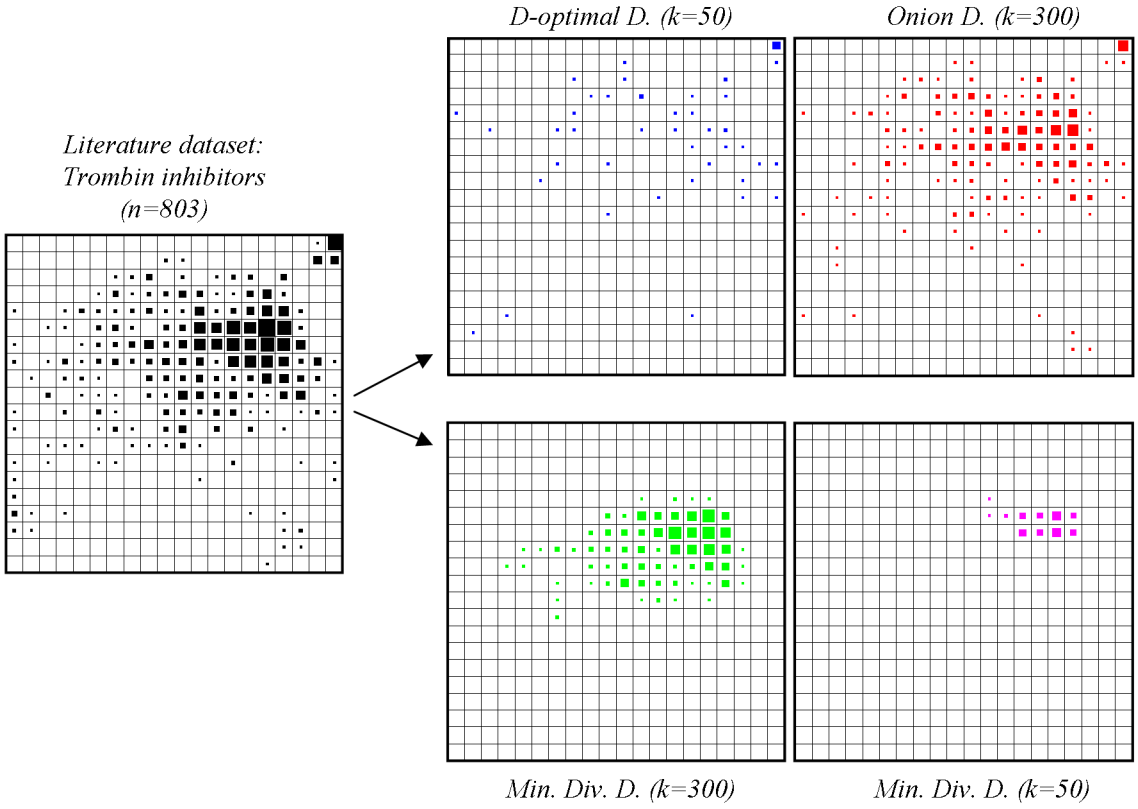


Figure 2.4: Visualization of the topology of sub-samples from the dataset of Thrombin inhibitors using Self-organizing maps (SOMs). Sub-samples are generated from the original dataset (black). The topology of the sub-samples varies according to sampling strategy and sample size (k). D-optimal Design with small k (blue) generates a set with the maximum degree of dispersion. Onion Design with large k (red) results in moderately patchy datasets with comprehensive coverage of the original dataset. Minimum diversity design with large (green) or small (magenta) k generates strongly patchy or concentrated datasets, respectively.

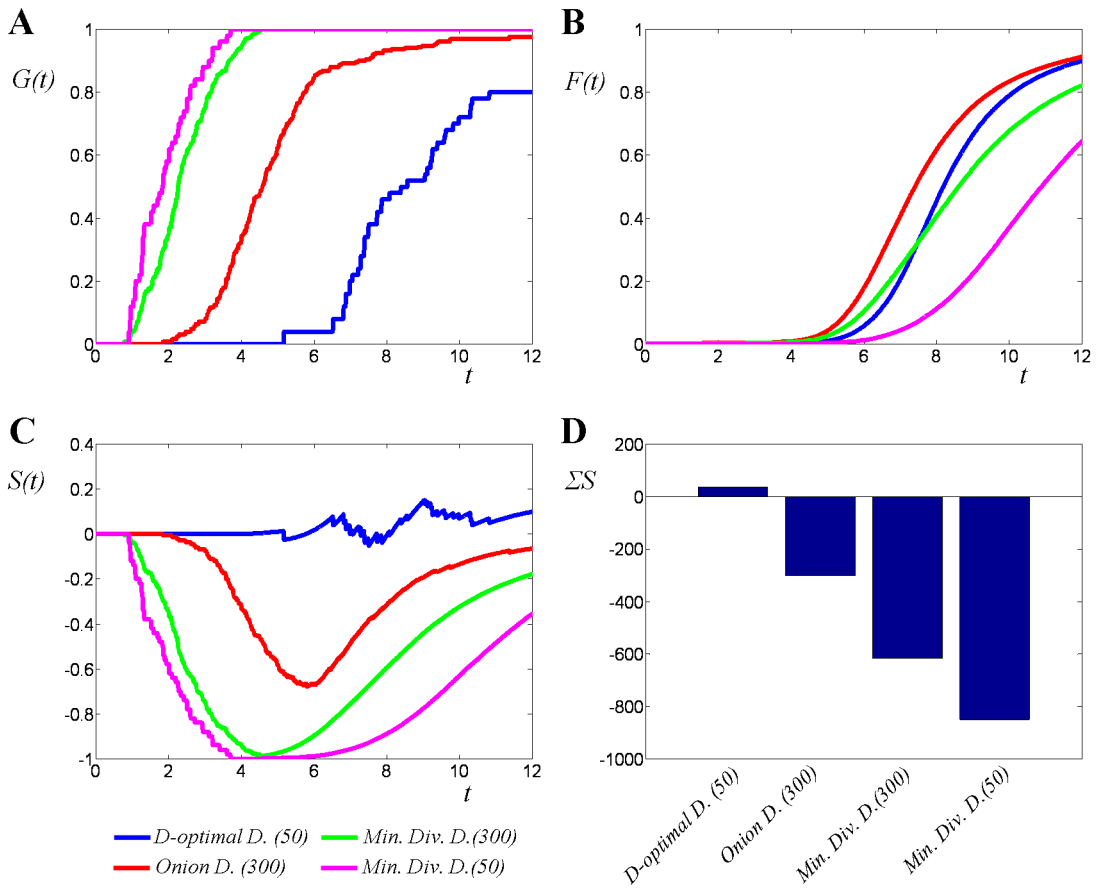


Figure 2.5: Topology analysis of four sub-samples from the dataset of Thrombin inhibitors. (A) Datasets from the Minimum diversity design (green, magenta) show the early and steep ascent in $G(t)$ characteristic for datasets with dense clumping. (B) The larger portion of chemical space occupied by the Onion design sub-sample (red) is indicated by the earlier and steep ascent in $F(t)$ as opposed to the concentrated sub-sample (magenta) from the Minimum Distance design. It is difficult to differentiate patchy (green) from dispersed (blue) datasets by $F(t)$ alone. (C) $S(t)$ facilitates an unambiguous characterization of sub-sample topology. The sub-samples are identified as dispersed (blue), moderately patchy (red), patchy with small separation (green) and concentrated (magenta). (D) ΣS reflects the dispersion of the D-optimal sample and the varying degree of clumping in the other samples.

Table 2.3: Correlation Coefficients of VS Performance Figures of Merit

	MOE	Simple
$\rho\,(mean(RTR_{1\%}), mean(AUC_{ROC}))^{a}$	0.92	0.85
Conf. Itv. Boundary^{b}	0.11	

[a] $n = 234$

[b] One-sided, 95%. Calculated using permutation testing. See Section 2.2.10.

descriptors. The resulting figures of merit as well as their variances and standard deviations ($mean(RTR_{1\%})$, $mean(AUC_{ROC})$, σ^2, σ^2_{top}, $\sigma^2_{bt,i}$, $std_{top}(FoM)$) are given in Table A.3 (Appendix). Although $mean(RTR_{1\%})$ and $mean(AUC_{ROC})$ are numerically different representations of VS performance, they agreed well in the relative rating of VS rankings for our experiments. In all the VS runs conducted here, there was no case where performance was rated high by $mean(RTR_{1\%})$ and intermediate or low by $mean(AUC_{ROC})$ and vice versa. The correlation coefficients of both figures of merit for MOE and simple descriptors are given by Table 2.3. All figures of statistical error were generally higher for RTR than for ROC. On the other hand, the discriminatory power of the RTR was found to be higher, particularly for the upper and lower ends of the VS performance spectrum. This kind of behavior is known and expected for these figures of merit and highlights again the advantages of their complementary use.

Analyzing the data shown in Tables A.1 and A.3, a strong correlation was detected between ΣS and VS performance. In general, a higher degree of clumping (indicated by large negative values of ΣS) accounts for better VS performance as measured by $mean(RTR_{1\%})$ and $mean(AUC_{ROC})$, for both MOE and simple descriptors. A summary of the observed correlation coefficients is given in Table 2.4 and an example for the relation of VS performance and sub-sample clumping is shown in Figure 2.6. Since negative values of ΣS indicate a higher degree of clumping, VS performance and ΣS are negatively correlated. Therefore, values of ρ close to -1 indicate a high degree of correlation between dataset clumping and VS performance.

ΣS only coarsely characterizes the degree of clumping in a dataset of compounds. The complex inter-relation of features like the number of clusters, their respective size, density and separation can not be reflected in detail by the simple scalar ΣS. In some pathological

Table 2.4: Correlation coefficients of VS performance figures of merit with ΣS.

	$\rho\,(\Sigma S, mean(RTR_{1\%}))^a$	$\rho\,(\Sigma S, mean(AUC_{ROC}))^a$
MOE	-0.93	-0.91
Simple	-0.89	-0.96
Conf. Itv. Boundary[b]	-0.11	

[a] $n = 234$

[b] One-sided, 95%. Calculated using permutation testing. See Section 2.2.10.

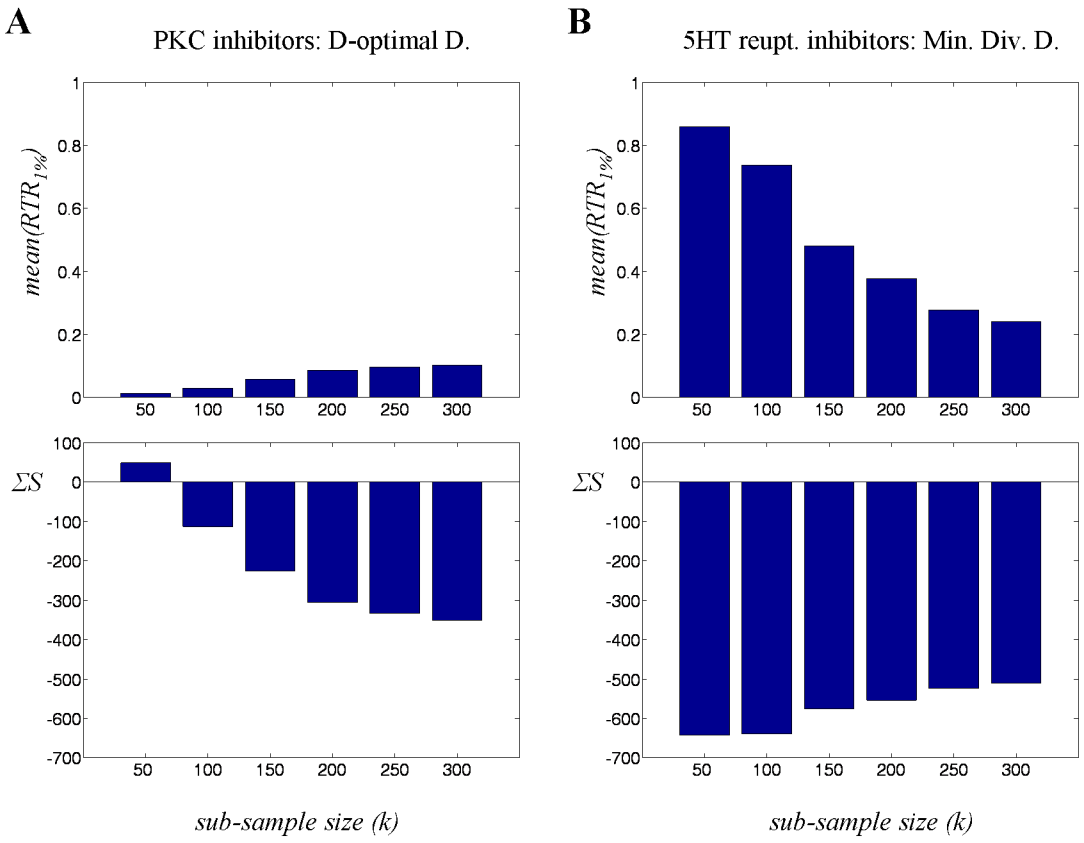


Figure 2.6: A strong correlation exists between $mean(RTR_{1\%})$ and ΣS, shown here for MOE descriptors. (A) Sub-samples generated by D-optimal Design from the dataset of PKC inhibitors show a small degree of clumping indicated by positive or small negative values of ΣS and low $mean(RTR_{1\%})$. (B) The high degree of clumping observed in Minimum Distance Design sub-samples of 5HT reuptake inhibitors is associated with high $mean(RTR_{1\%})$.

cases, the specific topology of two sub-samples can lead to very similar values of ΣS but to different outcomes of VS validation. This is illustrated by the extreme example of two sub-samples extracted from the dataset of Renin inhibitors (Onion D., $k = 50$) and 5HT reuptake inhibitors (Onion D.,$k = 300$). (Figure 2.7) Both samples have a similar ΣS of -454.4 and -453.8, but the $mean(RTR_{1\%})$ is 0.58 on the Renin inhibitors sub-sample and 0.18 on the sub-sample of 5HT reuptake inhibitors. When analyzing the graphs of $G(t)$ (Figure 2.7A), it is obvious that the average distances between actives in the sample of 5HT reuptake inhibitors are smaller, i.e. the density in the sample is higher. However, the respective graph for $F(t)$ (Figure 2.7B, red) with its prevalence of lower point-event distances shows quite clearly that this is mainly caused by small, local clusters evenly spread across chemical space. On the other hand, the sub-sample of Renin inhibitors is well separated from the rest of chemical space, a fact that dominates the validation result. (Figure 2.7B, blue) This is also well reflected in the graphs of $S(t)$, in which the separation of the Renin inhibitors from the background is indicated by a rightward shift. (Figure 2.7C) A SOM representation of both sub-samples (Figure 2.7D) provides an intuitive visualization of the respective conditions in the sub-samples. Summarizing, ΣS usually provides a robust and easily interpretable measure for dataset clumping and its effect on VS performance. However, to assess the topology in more detail, the graphs of $G(t)$, $F(t)$ and $S(t)$ have to be inspected. Here, visualization of dataset topology by Self-Organizing Maps can be of great benefit as it facilitates the intuitive perception of the information provided by the spatial statistics functions.

2.3.3 Topology Induced Component of Variance

The component of variance introduced by topology σ^2_{top} was found to be about 15 times larger (average over all sub-samples) in magnitude than the component associated with the experimental setup $mean(\sigma^2_{bt,i})$ for both, $RTR_{1\%}$ and AUC_{ROC}. Thus, sub-sample topology has a considerable impact on the statistical error of VS validations. However, it was not possible to deduce any relationship or correlation with the measures for dataset topology introduced here.

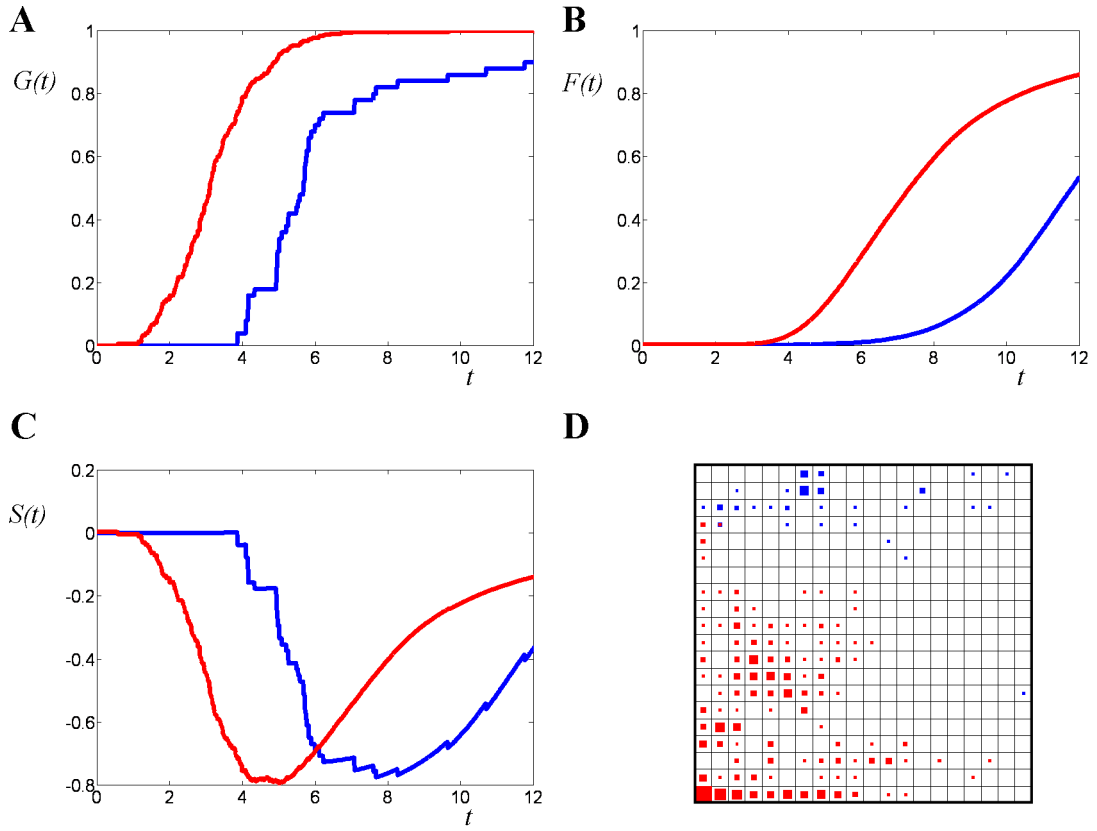


Figure 2.7: In some pathological cases estimation of clumping by ΣS may lead to ambiguous results. (A) The graph of $G(t)$ for the sub-sample of 5HT reuptake inhibitors (Onion D., $k = 300$; red) exhibits regions of higher density than the sample of Renin inhibitors (Min. Dist. D., $k = 50$, blue). (B) The smaller amount of empty space for the 5HT reuptake inhibitors sub-sample is evident from its graph for $F(t)$ (red). The larger amount of empty space present for the sample of Renin inhibitors (blue) indicates better separation of the sample from the background, causing better VS performance. (C) In this particular case, the graphs for $S(t)$ of both sub-samples are similar in shape and shifted on the x-Axis. Summing over $S(t)$ does not capture the rightward shift of the blue curve (Renin inhibitors) that indicates separation from the background. (D) The phenomena discussed in (A-C) are easily visualized on a SOM projection of the sub-samples. (Renin inhibitors: blue; 5HT reuptake inhibitors: red).

2.3.4 Comparison of Refined Nearest Neighbor Analysis with Other Approaches for Dataset Analysis

A commonly used measure to quantify dataset self-similarity in descriptor spaces is the average of intra-set pairwise distances (denoted as *avD*). In order to investigate if the characterization of dataset topology by Refined Nearest Neighbor Analysis really provides additional information over *avD*, the latter was computed for all sub-samples in MOE descriptor space. Since $S(t)$ and thus also ΣS combine a statistic of the self-similarity among actives in a dataset ($G(t)$, ΣG) and a statistic of the separation between decoys and actives in chemical space ($F(t)$, ΣF), the question arises whether any one of them can explain the differences in VS performance alone. It is also possible that other scalar values calculated from the cumulative distances functions $G(t)$ and $F(t)$, such as the respective medians, provide a more accurate characterization of dataset topology. Therefore the median nearest neighbor distance g_{med} was obtained as the value for t, where $G(t) = 0.5$. In an analogous fashion, the median distance f_{med} of a point to the nearest event was determined as the value of t where $F(t) = 0.5$, for each of the 234 sub-samples encoded by MOE descriptors. The values obtained for ΣG, ΣF, *avD*, g_{med} and f_{med} were correlated with VS performance on all sub-samples in the same way as ΣS. Additionally, the correlation coefficients were also determined separately for the different sub-sample design strategies (D-optimal D., Onion D., Min. Div. D.), in order to determine if any of the methods can capture dataset topology of a certain type (dispersed, patchy or concentrated) especially well.

Besides the already noted high correlation of $mean(RTR_{1\%})$ and ΣS, the results shown in Table 2.5 indicate moderate to high levels of correlation with $mean(RTR_{1\%})$ over all sub-samples of ΣG, *avD* and g_{med}, which are all based exclusively on active-active distances. The respective numbers for the different design strategies show however, that these levels of correlation for *avD* and g_{med} are mainly caused by sub-samples with concentrated topologies (Min. Div. D.). This result can be explained by the fact that for concentrated sub-samples, empty space will always be quite large with only marginal variations. Consequently, the main factor for their discrimination is the distribution of

Table 2.5: Correlation of several measures of dataset topology with $mean(RTR_{1\%})$.

$\rho\,(mean(RTR_{1\%}),...)$	ΣS	ΣG	ΣF	avD	g_{med}	f_{med}
All sub-samples[a]	-0.93	0.83	-0.61	-0.77	-0.53	0.44
Conf. Itv. Boundary[b]	-0.11	0.11	-0.11	-0.11	-0.11	0.11
D-optimal D.[b]	-0.92	0.81	0.23	-0.19	-0.07	0.31
Onion D.[b]	-0.74	0.27	-0.49	-0.04	-0.21	0.39
Min. Div. D.[b]	-0.95	0.76	-0.92	-0.79	-0.82	0.27
Conf. Itv. Boundary[c]	-0.19	0.19	-0.19	-0.19	-0.19	0.19

[a] $n = 234$

[b] $n = 78$

[c] One-sided, 95%. Calculated using permutation testing. See Section 2.2.10.

even-event distances, which is well captured by statistics like avD and g_{med}. However, this also shows, that avD and g_{med} can explain differences in VS performance only for concentrated datasets. Interestingly, ΣG retains a high level of correlation even for dispersed (D-optimal) sub-samples. Apparently, the characterization of the distribution of active-active distances by its numerical integral provides considerable additional information. A possible reason might be, that g_{med} only captures the shift but not the slope of the ascent of $G(t)$, which can indicate the presence of small local clusters. (See Section 2.2.8.2)

On the other hand, ΣF captures considerable information for concentrated (Min. Div. D.) and patchy (Onion D.) sub-samples. If a sub-sample is located in a region of chemical space, which is sparsely populated by decoys (i.e. more empty space, small ΣF) VS performance will increase. However, ΣF seems unable to capture this relation on D-optimal sub-samples, whereas f_{med} shows low, but constant levels of correlation for all sub-samples.

Summarizing, all of these measures are able to reflect the impact of dataset topology on VS performance under certain conditions, but only ΣS provides a comprehensive coverage across all types of dataset topology, as indicated by the consistently higher correlation coefficients for ΣS in Table 2.5. Put another way, neither statistics about intra-set distances, i.e. dataset self-similarity, nor information about empty space, i.e. separation between decoys and actives, can explain differences in VS performance alone. The augmentation of the nearest-neighbor function $G(t)$ by the empty space function $F(t)$

provides a substantial amount of additional information.

2.3.5 Mapping Performance

It was shown above, that the success rate of virtual screening is higher on datasets that have a clumpy topology in descriptor space. In a way, this is not surprising, since this is exactly what chemical descriptors were invented for. Good descriptors map compounds with similar bioactivities to similar points in descriptor space, thereby introducing clumping. However, some descriptors are better at this task and some are worse. Put differently: If the same dataset (or in our case a sub-sample of a dataset) is encoded by two different descriptors, VS performance should be better using the descriptor representation that introduces the more favorable topology, i.e. more clumping. (Figure 2.8) Thus, the *"mapping performance"* of one descriptor vs. another can be quantified by the additional degree of clumping introduced by the former descriptor. Applied to the two descriptors used in this chapter, the mapping performance of MOE descriptors vs. simple descriptors is given by the difference in ΣS on a given dataset or sub-sample:

$$\Delta(\Sigma S) = \Sigma S_{MOE} - \Sigma S_{simple}; \tag{2.14}$$

with $\Delta(\Sigma S)$ the difference in clumping, i.e. the mapping performance. Since a higher degree of clumping is associated with negative values of ΣS, negative values of $\Delta(\Sigma S)$ indicate more clumping for the representation by MOE descriptors, i.e. a higher mapping performance of MOE vs. simple descriptors.

For the two descriptors (MOE and simple) used here, the difference in VS performance was calculated for each sample as:

$$\Delta(FoM) = mean(FoM_{MOE}) - mean(FoM_{simple});\ FoM = RTR_{1\%}\ or\ AUC_{ROC} \tag{2.15}$$

for both RTR and ROC, respectively. Accordingly, $\Delta(FoM)$ is positive, whenever MOE performs better and negative if the simple descriptors generate superior figures of

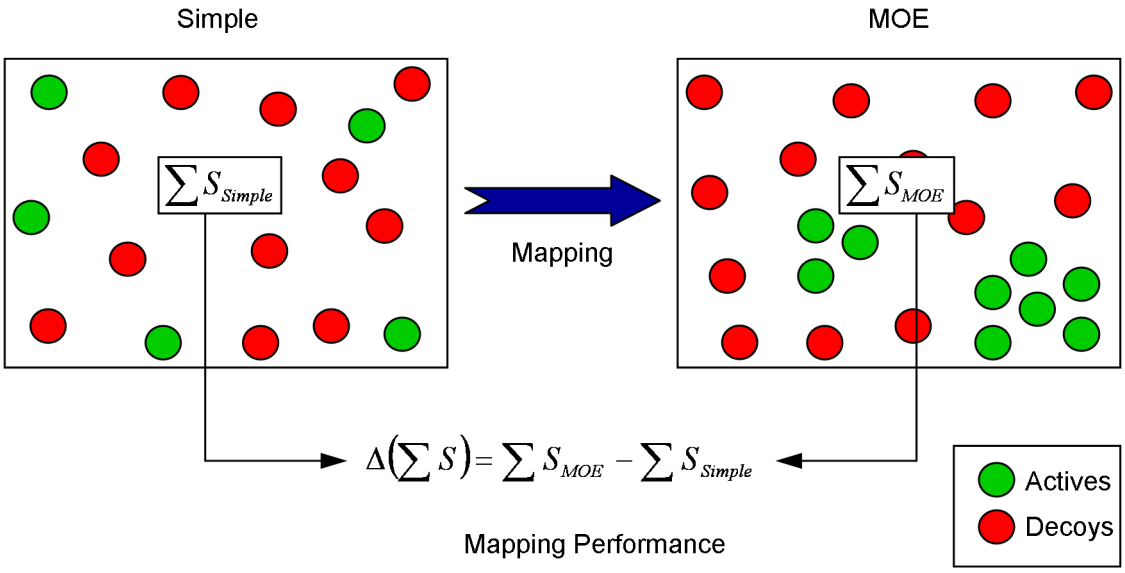


Figure 2.8: A descriptor that maps a dataset to a more clumpy topology than another descriptor will generate better results in VS validation experiments. The mapping performance of a descriptor vs. another descriptor is given by the difference of the respective values of ΣS on a given dataset $\Delta(\Sigma S)$.

Table 2.6: Correlation of the difference in VS performance with the mapping performance of MOE vs. simple descriptors.

	$\Delta(RTR_{1\%})$	$\Delta(AUC_{ROC})$
$\rho(\Delta(\Sigma S), \ldots$[a]	-0.93	-0.88
Conf. Itv. Boundary[b]		-0.11

[a] $n = 234$

[b] One-sided, 95%. Calculated using permutation testing. See Section 2.2.10.

merit.

For all sub-samples examined in this chapter, a strong correlation was found between $\Delta(\Sigma S)$ and $\Delta(FoM)$. The respective spearman correlation coefficients are given in Table 2.6.

An example for sub-samples taken from the ACE inhibitors dataset is given in Figure 2.9. Better VS performance for a certain type of descriptors is closely associated by a better mapping performance on the respective sub-sample. It should be recalled, that the D-optimal Design sub-samples with $k = 50$ constitute an artificial absolute "worst-case" scenario for performing VS with MOE descriptors. Thus, since it cannot get any worse, it is not surprising that the representation of the same compounds in simple descriptor space is more clumpy. Therefore both, the better mapping performance and the better

VS performance of simple descriptors on these sub-samples are mainly an effect of the sampling strategy.

2.3.6 Application to Whole Datasets

Using sub-samples generated by different design strategies, it was possible to observe the impact of dataset topology on the results of VS validation. Furthermore, it was shown, that this impact can be quantified using spatial statistics methods. However, when performing a real-life validation of a ligand-based VS technique, one wouldn't be interested in the VS performance on artificial sub-samples, but on the complete benchmark datasets.

Table 2.7 and Figure 2.10 summarize the results of retrospective VS simulations following the procedure stated above (100 query / validation set splits, 10 query compounds, MAX-rule data fusion) and Refined Nearest Neighbor Analysis on the complete benchmark datasets. Again, a strong correlation between dataset clumping and VS performance as measured by ΣS, $mean(RTR_{1\%})$ and $mean(AUC_{ROC})$, respectively, can be observed (Table 2.7). The somewhat smaller values for the correlation coefficients are mainly due to the much smaller number of samples (13 complete datasets vs. 234 sub-samples) available for their calculation, so that deviations from the generally observed relationship have a higher impact. Once more, it must be stated that ΣS is a robust but rough estimate of dataset clumping, which can not explain all effects of dataset topology on VS performance perfectly. Discrepancies from the overall correlation of ΣS and VS performance such as those observed in Figure 2.10, can usually be explained by a more detailed inspection of $G(t)$, $F(t)$ and $S(t)$.

Also, the correlation of mapping performance and VS performance persists for the complete datasets. (Figure 2.11, Table 2.9). Apart from minor discrepancies, $\Delta(\Sigma S)$ is able to explain most of the differences in VS performance. Large gains in clumping are always associated with much better performance of the respective descriptor and small differences in topology coincide with small or no changes in VS performance. For answering the question if a particular descriptor really improves VS performance in a real-life VS campaign these more general trends are of higher importance.

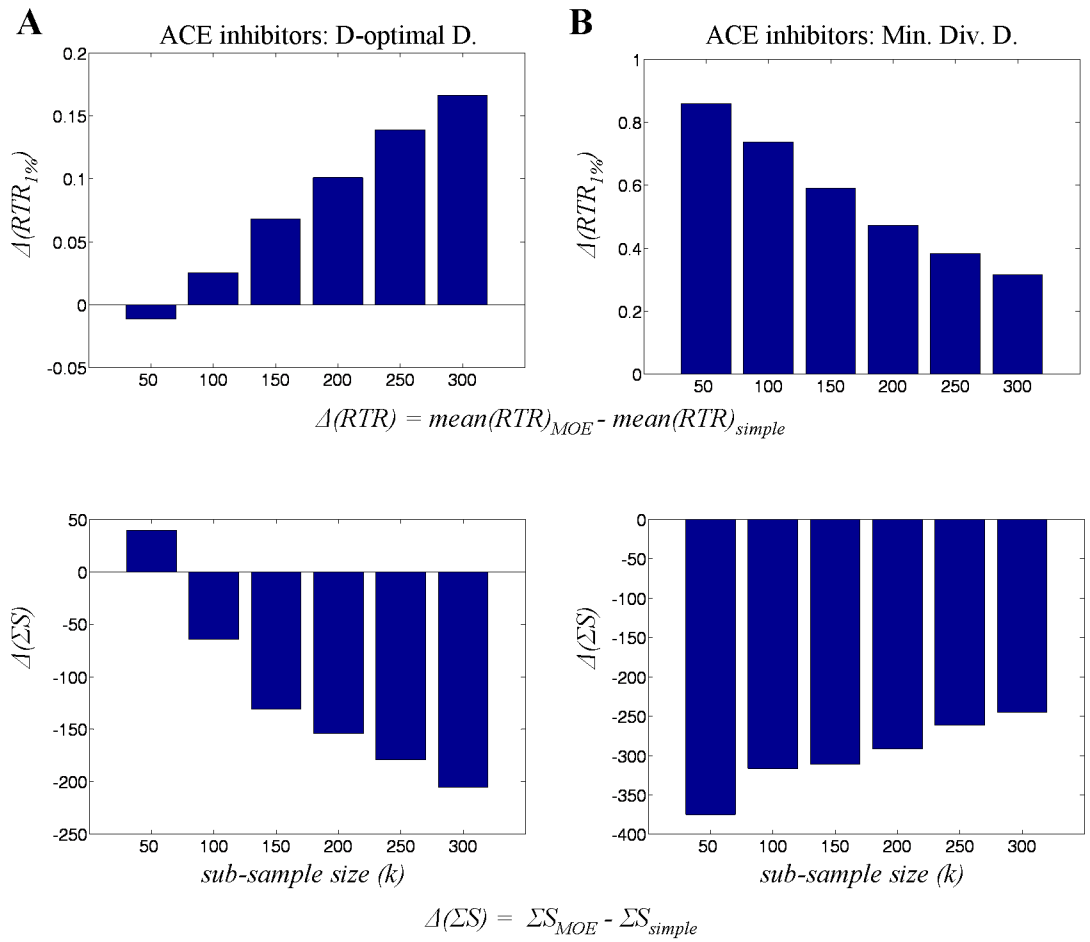


Figure 2.9: Mapping performance of MOE descriptors vs. simple descriptors for sub-samples of the ACE inhibitors dataset generated by D-optimal Design (A) and Minimum Diversity Design (B), respectively. (A) $\Delta(\Sigma S)$ is positive for $k = 50$, meaning that the representation of the sub-sample by simple descriptors features a higher degree of clumping. Accordingly, $\Delta(RTR_{1\%})$ is negative for that sub-sample, indicating the better performance of simple descriptors. For $k > 50$, $\Delta(\Sigma S)$ shows negative values and is associated with positive values for $\Delta(RTR_{1\%})$, which indicates that MOE outperforms simple with respect to $mean(RTR_{1\%})$ and clumping. This is due to the fact, that the D-Optimum Design sub-samples get clumpier with increasing k, since more datapoints from the bulk of the dataset have to be selected. (B) The Minimum Diversity Design constitutes the "best-case" scenario for MOE descriptors. Accordingly both, $\Delta(RTR_{1\%})$ and $\Delta(\Sigma S)$ indicate higher performance of MOE-PCA descriptors in these sub-samples. Here, the effect of increasing k is inverse to the D-Optimum Design. Starting with the very set of minimally diverse compounds at $k = 50$, more and more compounds from the outer regions of the dataset have to be selected for larger k, rendering the sub-sample more diverse.

Table 2.7: VS performance and degree of dataset clumping for complete benchmark datasets

	MOE			Simple		
	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$	ΣS	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$	ΣS
ACE inhibitors	0.34	0.79	-541.5	0.10	0.65	-316.7
AChE Inhibitors	0.17	0.81	-455.4	0.09	0.77	-386.4
Angio. R. Blockers	0.23	0.84	-474.8	0.15	0.84	-445.8
COX inhibitors	0.11	0.78	-387.4	0.08	0.82	-387.5
D2 antagonists	0.19	0.87	-457.3	0.11	0.83	-411.6
HIV P. inhibitors	0.16	0.78	-454.7	0.09	0.73	-352.5
5HT1A agonists	0.16	0.87	-484.1	0.11	0.82	-421.1
5HT3 antagonists	0.29	0.92	-507.5	0.26	0.92	-495.5
5HT reup. inhibitors	0.20	0.86	-488.6	0.15	0.83	-433.4
PKC inhibitors	0.19	0.69	-409.6	0.07	0.57	-281.3
Renin inhibitors	0.45	0.93	-602.8	0.50	0.96	-692.2
Subst. P inhibitors	0.16	0.79	-438.2	0.07	0.73	-367.3
Thrombin inhibitors	0.23	0.82	-443.6	0.13	0.76	-342.8

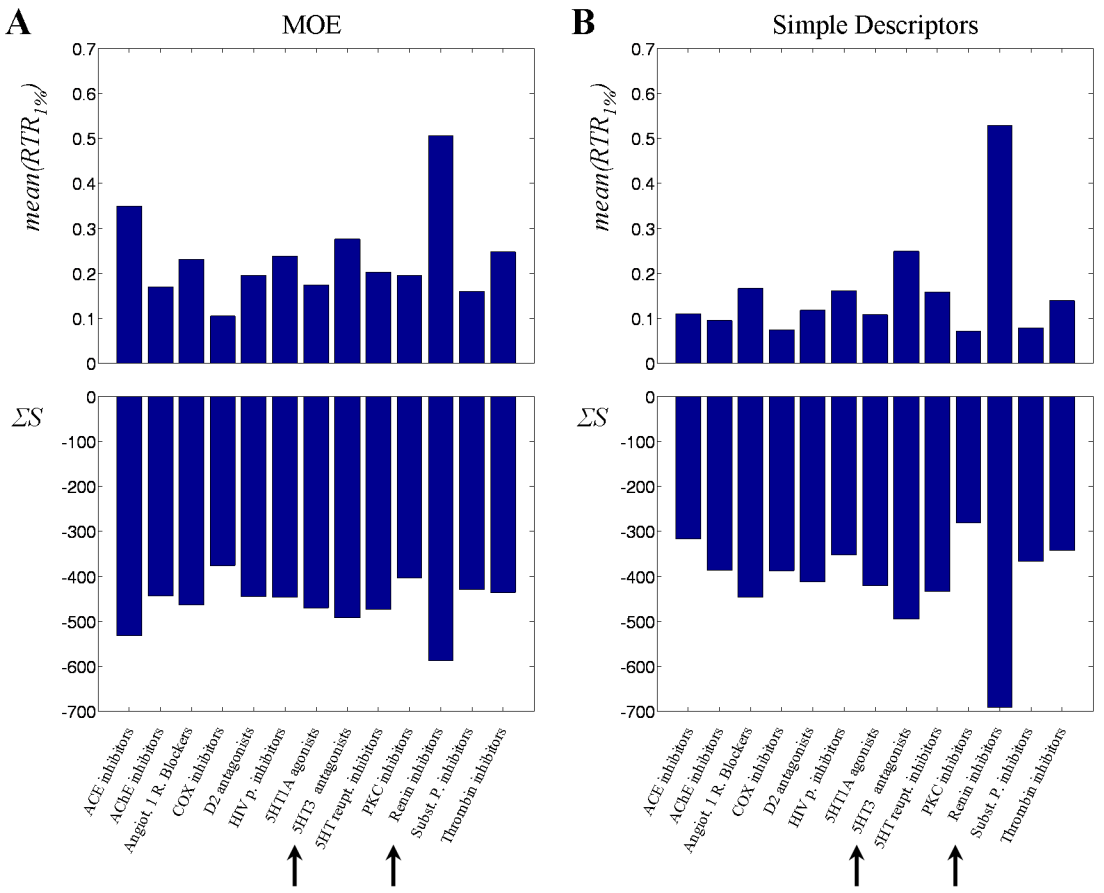


Figure 2.10: VS performance and dataset clumping as observed on the complete benchmark datasets. As shown for the sub-samples of controlled topology, a strong correlation exists between dataset clumping and VS performance. Arrows indicate datasets discussed in more detail in the text.

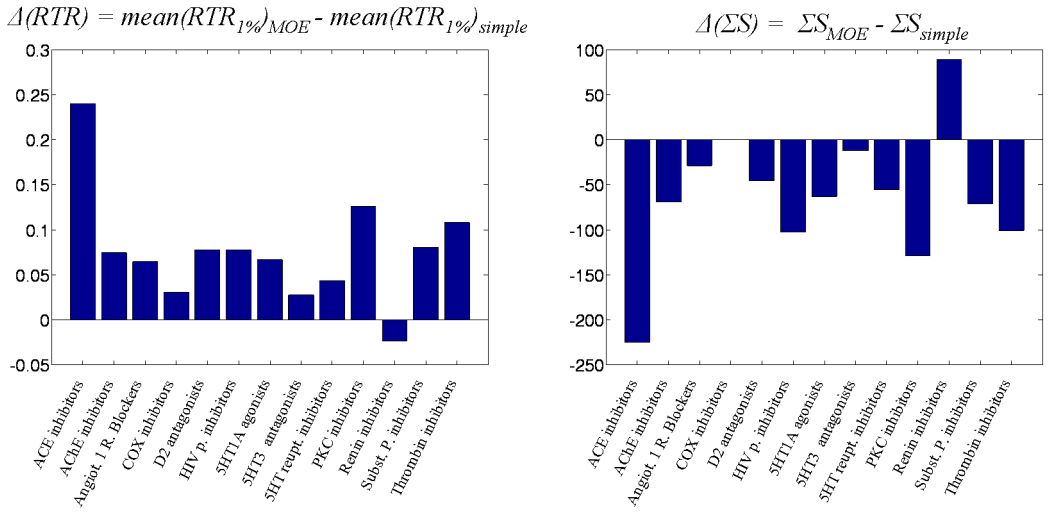


Figure 2.11: Differences in VS performance and gain in clumping of MOE vs. simple descriptors on the complete benchmark datasets. A strong overall correlation of $\Delta(\Sigma S)$ and differences in VS performance can be observed.

Table 2.8: Correlation of VS Figures of Merit with Dataset Clumping (ΣS) for the Complete Benchmark Datasets.

	$\rho\,(\Sigma S, mean(RTR_{1\%}))^a$	$\rho\,(\Sigma S, mean(AUC_{ROC}))^a$
MOE	-0.72	-0.77
simple	-0.77	-0.97
Conf. Itv. Boundary[b]	-0.48	

[a] $n = 13$

[b] One-sided, 95%. Calculated using permutation testing. See Section 2.2.10.

Table 2.9: Correlation of the difference in VS performance with the mapping performance of MOE vs. simple descriptors on whole datasets.

	$\Delta(RTR_{1\%})$	$\Delta(AUC_{ROC})$
$\rho(\Delta(\Sigma S), ...^a$	-0.80	-0.94
Conf. Itv. Boundary[b]	-0.48	

[a] $n = 13$

[b] One-sided, 95%. Calculated using permutation testing. See Section 2.2.10.

2.3.7 Benchmark Dataset Bias

From the chart of ΣS for the benchmark datasets encoded by the simple descriptors (Figure 2.10), the high degree of clumping observed for the datasets of Renin inhibitors and 5HT3 antagonists is striking. The values of ΣS for these datasets show, that they can easily be distinguished from the background even by these absolutely simple descriptors that do not encode molecular connectivity. Accordingly, it should not be difficult to achieve good results when performing VS validation on these datasets. So, these datasets introduce a bias towards good validation results, which is also reflected by the fact that the highly complex MOE descriptors can not generate any significant gain in clumping over simple descriptors on these datasets. Actually both, clumping and VS performance are lower for the Renin inhibitors dataset with MOE than with simple descriptors. Thus, the $mean(RTR_{1\%})$'s of ~ 0.3 and ~ 0.5 achieved with MOE on these datasets, which would normally be considered quite acceptable, are of no real value for the evaluation of the VS capabilities of MOE descriptors. We term this tendency towards overoptimistic validation results "benchmark dataset bias". The results presented here show, that for ligand-based virtual screening "benchmark dataset bias" does not only depend on the distances between

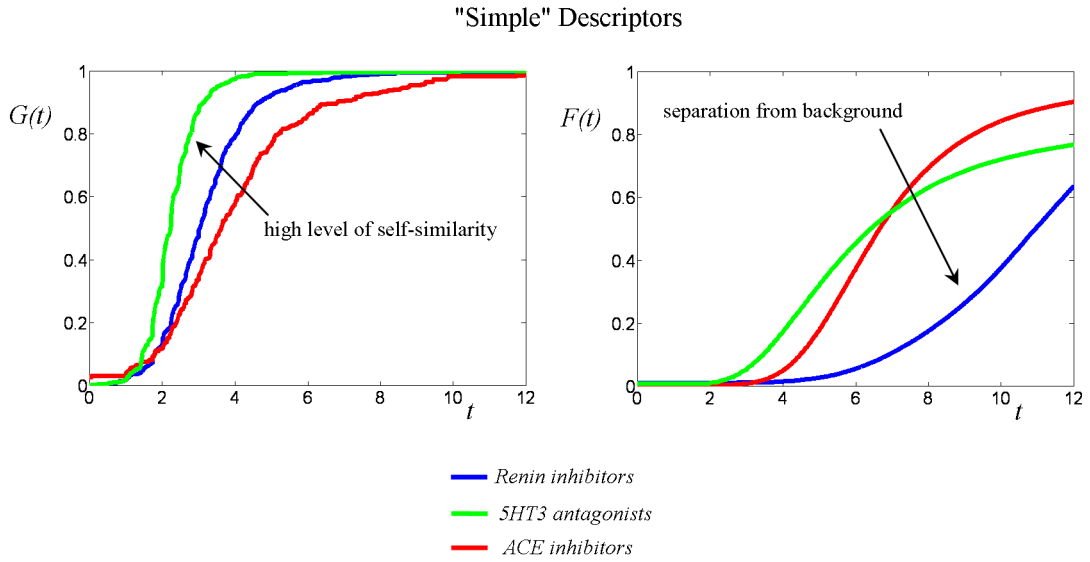


Figure 2.12: Benchmark dataset bias is caused by both, high levels of intra-set self-similarity in the datasets of actives and separation of actives from the background. Renin inhibitors (blue) and 5HT3 antagonists (green) both show benchmark dataset bias caused by a high degree of clumping in simple descriptor space. A comparison of the respective graphs for $G(t)$ and $F(t)$ shows, that for the 5HT3 antagonists this is mainly caused by small active-active (i.e. event-event) distances. On the Renin dataset, it is mainly due to a high degree of separation from the rest of chemical space. In comparison, the dataset of ACE inhibitors (red), which is not subject to benchmark dataset bias, exhibits neither self-similarity nor separation from the background.

actives themselves (i.e. $G(t)$), but also on the separation between actives and inactives (i.e. $F(t)$), a fact that should be taken into account in the design of validation experiments comparing ligand based virtual screening methods (Figure 2.12). As a first consequence, datasets that show a high degree of clumping in a simple reference descriptor space should be avoided in VS validation experiments. If, for some reason they cannot be evaded, the value of ΣS determined in a simple reference descriptor space can effectively be used to estimate an expectation for VS performance. Moreover, using the information provided by $F(t)$ and $G(t)$, it is possible to determine if the clumping in simple descriptor space is mainly caused by a high degree of separation of the dataset from the background, e.g. the dataset of Renin inhibitors (Figure 2.12). In this case, the problem can be solved simply by choosing a more appropriate decoy dataset.

2.4 Summary

The work presented in this Chapter introduces spatial statistics methods to the field of chemoinformatics. Thus far, Refined Nearest Neighbor Analysis has been mostly applied to problems from the sciences of ecology or forestry, in which spatial data can usually be represented by two-dimensional, finite maps.[93,94] The boostrapping procedure developed in Section 2.2.8.3 allows the application of Refined Nearest Neighbor methods to high-dimensional, non-finite chemical spaces. Utilizing this adapted form of Refined Nearest Neighbor Analysis, a framework was developed that provides tools for the topological analysis of chemical datasets in all degrees of detail, ranging from a quick estimate of clumping to an in-depth analysis of clustering behavior. The cumulative distribution functions describing the spatial relations between actives and decoys in a VS benchmark dataset were employed to derive quantitative, scalar measures for both analogue bias (ΣG) and artificial enrichment (ΣF). Furthermore, a combined measure of dataset clumping (ΣS) comprising both, analogue bias and artificial enrichment, was introduced.

By correlation studies based on the results of topological analysis and retrospective VS simulations carried out on dataset sub-samples with defined topology it was shown, that the topology of benchmark datasets in descriptor space has a considerable impact on the results of VS validation. These results point out that in contrast to molecular docking, both, the mutual distances of the active compounds (i.e analogue bias) and their separation from the decoys (i.e. artificial enrichment) in descriptor space are of importance for the validation of ligand-based virtual screening techniques. Furthermore, the methodology proposed in this Chapter provides insights about the reasons for differences in the VS performance of different descriptors. It was shown, that better VS performance of a descriptor can be linked to a superior mapping performance.

The methodology for the characterization of dataset topology presented here does not imply any prior assumptions or preconditions about the composition of the datasets. On the contrary, it is actually suited to provide exactly this information, i.e. if the dataset is composed of a single or multiple clusters and if these are close or separated in chemical space.

An obvious field of potential future use for this piece of information would be the rational design of validation experiments comparing different algorithms for similarity searching. On a patchy or dispersed dataset, for instance, any algorithm based on multiple query molecules should be superior to an algorithm with only one query as an input. This advantage would be annihilated on a concentrated dataset. In this context, an unbiased selection of benchmark datasets could be greatly facilitated by the topology analysis proposed here.

The methodology uses two basic functions for the elucidation of benchmark dataset topology: The nearest-neighbor function $G(t)$ reflects the distribution of intra-set "active-to-active" distances, whereas the empty-space function $F(t)$ represents the distribution of "decoy-to-active" distances. As shown in Section 2.3.7, bias introduced to VS validation experiments by the composition of benchmark datasets can be detected and quantified by these functions. This opens up another field of application for the methodology: the design of validation datasets that minimize dataset induced bias. The results presented in this Chapter show that clumpy topologies in simple descriptor space are the cause for the occurrence of analogue bias and artificial enrichment. Consequently, benchmark dataset bias can be prevented by designing datasets with non-clumpy topologies regarding simple descriptors. Chapter 3 will present a workflow that applies this rationale to the design of unbiased datasets for LBVS benchmarking based on bioactivity datasets extracted from PubChem.

Chapter 3

Maximum Unbiased Validation (MUV) Datasets

3.1 Objectives

The results obtained from the investigations described in Chapter 2 will be utilized to derive criteria for the design of validation datasets without benchmark dataset bias. Based on these criteria, objective functions will be formulated and and a workflow will be devised for the rational design of Maximum Unbiased Validation (MUV) datasets. This procedure will be applied to a collection of bioactivity data extracted from PubChem in order to design publicly available unbiased datasets for benchmarking. The origin of PubChem bioactivity data from high throughput screening (HTS) experiments and the associated error rates (see Section 1.1.2) make it necessary to scrutinize the reported bioactivities with extreme thoroughness. A data centered workflow will be developed that purges the bioactivity datasets from compounds for which the respective activity may be subject to any doubts.

3.2 Methods

3.2.1 Criteria for Refined Nearest Neighbor Analysis Based Benchmark Dataset Design

In Chapter 2 it was shown that the validation of LBVS methods is affected by both, artificial enrichment and analogue bias. Indeed, datasets usually exhibit a combination of both phenomena. Refined Nearest Neighbor Analysis, a non-parametric methodology from the framework of spatial statistics, was used to quantify the effect of benchmark dataset bias by an analysis of the dataset's topology in chemical space. Refined Nearest Neighbor Analysis provides the figures ΣS, ΣG and ΣF, which estimate "dataset clumping", self-similarity in the dataset of actives and separation between decoys and actives, respectively. In particular, negative values of ΣS indicate dataset clumping, positive values indicate dispersion and values near zero a spatially random distribution of actives and decoys. It was demonstrated in Chapter 2 that over-optimistic VS validation results and benchmark dataset bias can be linked to dataset clumping.

For the design of unbiased benchmark datasets that are not affected by artificial enrichment and analogue bias, ΣG and ΣF can effectively be employed as objective functions. However, Refined Nearest Neighbor Analysis is based on a representation of datasets by descriptors. Hence, for the calculation of ΣG and ΣF a descriptor must be utilized that captures the molecular properties associated with benchmark dataset bias. For this purpose, the "simple" descriptors introduced in Section 2.2.3 are perfectly suited, since they have been shown in Section 2.3.7 to capture both, analogue bias and artificial enrichment. As a first condition for unbiased validation experiments, datasets should always exhibit a dispersed topology in the chemical space spanned by these simple descriptors. More precisely, active-active distances should be larger than or equal to decoy-active distances in simple descriptor space, because otherwise similarity searching becomes trivial. Topologies of this type have been shown to prevent both, analogue bias and artificial enrichment in Chapter 2.

Apart from these more general conditions for validation dataset design, it is important

to keep in mind the existence of two basic experimental settings for VS validation as discussed in Section 1.2.1.

Suitability testing is used to optimize a VS protocol for a specific target or target class. As a consequence, datasets for suitability testing must cover the respective activity space comprehensively. Differences between the activity spaces of different targets must be truly reflected by suitability testing datasets.

The objective of this Chapter will be the design of datasets for *benchmarking experiments.* In contrast to suitability testing, the goal of benchmarking experiments is to identify the method with the best VS performance across a range of datasets. In order to maximize the information of such experiments regarding the potential of different methods, it is desirable to minimize the influence of dataset topology on validation results. Hence, the *differences* in dataset composition must be minimized, i.e. all datasets should be adjusted to a common level of dispersion in simple molecular property space. Any arbitrary level of dispersion would suffice for benchmarking experiments of this type, as long as it is common to all datasets. However, the state of "spatial randomness" ($\Sigma S \approx 0$) is especially advantageous, since it implies that the distribution of simple molecular properties between actives and decoys contains no information about the respective bioactivities. The benchmark datasets presented here were therefore designed to be as close to spatial randomness as possible given the topology of the original datasets.

As stated above, datasets with a spatially random topology in simple descriptor space are unbiased with respect to analogue bias and artificial enrichment. Therefore, they constitute a tool for the Maximum Unbiased Validation (MUV) of virtual screening techniques.

3.2.2 MUV Benchmark Dataset Design Strategy

The design strategy for MUV datasets comprised three major steps (Figure 3.1):

(1) A collection of bioassays was extracted from PCBioAssay that justifies high confidence in the respective bioactivities. The compounds found active and inactive, respectively, formed the basis for the subsequent design steps. The resulting datasets of com-

pounds were termed "potential actives" (*PA*) and "potential decoys" (*PD*). Filters were applied to the *PA* datasets that further purged all compounds for which the specificity of the respective bioactivities might be subject to any doubts.

(2) The chemical space around the *PA* compounds was examined statistically in order to determine if the *PA* compounds are well embedded in decoys, a precondition for validation set design. (3) Experimental design algorithms were applied to select subsets of $k = 30$ actives and $d = 15000$ decoys from the *PA/PD* datasets with a spatially random distribution of actives and decoys regarding simple molecular properties. With constant dataset sizes of $k = 30$ and $d = 15000$, MUV datasets also minimize the variance of validation results as demonstrated by Truchon et al.[60] The numerical values of k and d were chosen arbitrarily based on the size of the available bioactivity datasets. In principle, k and d could also be set to other values.

3.2.3 PubChem as a Source of VS Validation Datasets

As stated above (Section 1.1.2.2), PubChem[15] is the central repository of small molecule data of the Molecular Libraries Initiative[12] of the NIH Roadmap for Medical Research[13,14] and is composed of three major databases. Two will be used here: PCCompound, which provides chemical structures of the tested compounds, and PCBioAssay, which lists the respective bioactivity data. Each record in PubChem is assigned a unique ID (UID) by which it can be easily accessed and retrieved. For PCCompound and PCBioAssay, these IDs are termed compound ID (CID) and assay ID (AID), respectively. Compared to other databases of bioactivity data, PubChem features several major advantages with respect to the design of VS benchmark datasets:

(i) All data in PubChem, including structures of compounds, bioassay conditions and experimental readouts are publicly accessible. (ii) Due to the specifications of the NIH Roadmap initiative, the compound collections tested in each bioassay exhibit a remarkable level of diversity. (iii) The vast majority of tested compounds are "drug-like". (iv) For each assay, compounds that were found to be inactive are listed in addition to those found to be active. These inactive compounds can be used as decoys in validation ex-

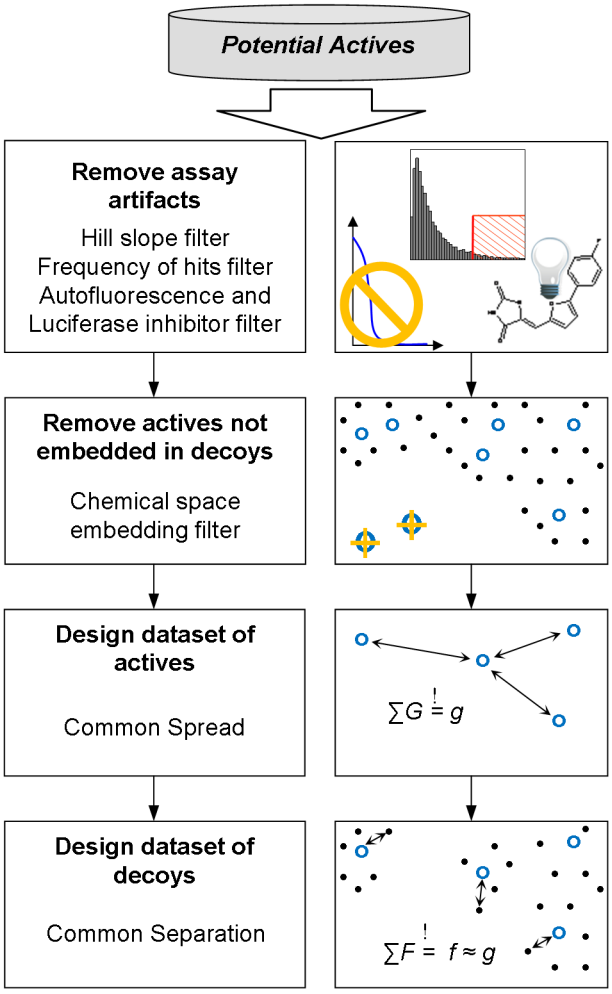


Figure 3.1: Synopsis of the MUV design workflow. Compounds with a potential for unspecific bioactivity are removed by the Assay artifacts filter. Actives devoid of decoys are removed by the Chemical space embedding filter. The spread of actives (ΣG) is adjusted to a common level g. Correspondingly, the separation between actives and decoys (ΣF) is adjusted to a common level f that is similar to g, thereby enforcing approximate spatial randomness.

periments, as done recently by Hsieh et al. for a single dataset in the validation of a QSAR modeling approach.[46] This provides the unique opportunity to design decoy sets, for which the inactivity against the target is actually experimentally validated. (v) PubChem is fully integrated into the NCBI Entrez database system.[16] Using the Entrez Programming Utilities (E-Utilities)[16] and the PubChem Power User Gateway (PUG)[105] automated chemogenomics analyses are feasible, linking compounds with their bio-activities and the protein or DNA information of their targets.[106] On the downside however, most of the bioactivity data available from PubChem is based on High-Throughput Screening (HTS) experiments. HTS data is notoriously affected by experimental noise and artifacts.[11,17,44,107] Thus, for the design of benchmark datasets it is essential to scrutinize PubChem bioactivity data with extreme thoroughness.

3.2.4 Selection of Bioactivity Datasets

All assays with a specified protein target were extracted from PCBioAssay. From these, pairs of primary and confirmatory bioassays against the same target were selected. In these pairs, the bioactivity against the target is first determined for a large set (>50000) of compounds in a primary HTS experiment. The hits from the primary screen are then subjected to a low-throughput confirmatory screen testing for dose-response relationships. (see Section 1.1.2) To be selected for the MUV design process, the confirmatory screens were further required to contain associated EC_{50} values. The actives from the confirmatory screens, referred to as Potential Actives (*PA*) in the text, and the inactives from the primary screens, referred to as Potential Decoys (*PD*), formed the basis for the generation of MUV datasets as presented here. The resulting datasets are summarized in Table 3.1. SD-files of the datasets were downloaded from PubChem by a Perl script that utilizes the PubChem Power User Gateway.[105]

3.2.5 Assay Artifacts Filter

The requirement, that the bioactivity of all potential actives is determined by low-throughput dose-response experiments, justifies some confidence in the reliability of these assign-

Table 3.1: Bioactivity Datasets from Pairs of Primary HTS and Confirmatory Dose-response Experiments.

Target	Mode of Interaction	Target Class	Prim. Assay (AID)	Confirm. Assay (AID)	Assay-Type	Hill Slope[a]	PA	PD
S1P1 rec.	Agonists	GPCR	449	466[b]	Reporter Gene	PubChem	223	55395
PKA	Inhibitors	Kinase	524	548	Enzyme	PubChem	62	64814
SF1	Inhibitors	Nuclear Receptor	525	600	Reporter Gene	PubChem	213	64550
Rho-Kinase2	Inhibitors	Kinase	604	644	Enzyme	PubChem	67	59576
HIV RT-RNase	Inhibitors	RNase	565	652	Enzyme	PubChem	370	63969
Eph rec. A4	Inhibitors	Rec. Tyr. Kinase	689	689[c]	Enzyme	PubChem	80	61480
SF1	Agonists	Nuclear Receptor	522	692	Reporter Gene	PubChem	75	63683
HSP 90	Inhibitors	Chaperone	429	712	Enzyme	Calculated	91	63481
ER-α-Coact. Binding	Inhibitors	PPI	629	713	Enzyme	Calculated	221	84656
ER-β-Coact. Binding	Inhibitors	PPI	633	733	Enzyme	Calculated	194	84984
ER-α-Coact. Binding	Potentiators	PPI	639	737	Enzyme	Not applicable[d]	64	84947
FAK	Inhibitors	Kinase	727	810	Enzyme	PubChem	110	96070
Cathepsin G	Inhibitors	Protease	581	832	Enzyme	PubChem	65	62007
FXIa	Inhibitors	Protease	798	84	Enzyme	PubChem	70	218421
S1P2 rec.	Inhibitors	GPCR	736	851	Reporter Gene	PubChem	54	96674
FXIIa	Inhibitors	Protease	800	852	Enzyme	PubChem	99	216795
D1 Rec.	Allosteric Modulators	GPCR	641	858	Reporter Gene	PubChem	226	54292
M1 Rec.	Allosteric Inhibitors	GPCR	628	859	Reporter Gene	PubChem	231	61477

[a)] Large Hill slopes are generally considered harbingers of unspecific inhibition. See Section 3.2.5.1 for details.

[b)] A counter-screen was conducted using negative control cells (AID: 467). Actives as reported here are compounds active in assay 466 but not in assay 467.

[c)] Primary high-throughput assay and confirmatory dose response assay were carried out, but reported as one record in PCBioAssay.

[d)] The mechanism of the PPI potentiation does not comply with Michaelis-Menten kinetics

ments. Nevertheless, screening experiments are prone to artifacts caused by the tendency of some organic chemicals to form aggregates in aqueous buffers,[11] to exert off-target or cytotoxic effects[108] or to interfere with the assay's optical detection method.[109,110] In order to remove all compounds for which the specific mode of action could be subject to doubts from the PA datasets, a range of filters was applied, namely the "Hill slope filter", "Frequency of hits filter" and the "Autofluorescence and Luciferase inhibition filter". Together these filters form the "Assay artifacts filter". The particular filters were implemented as follows.

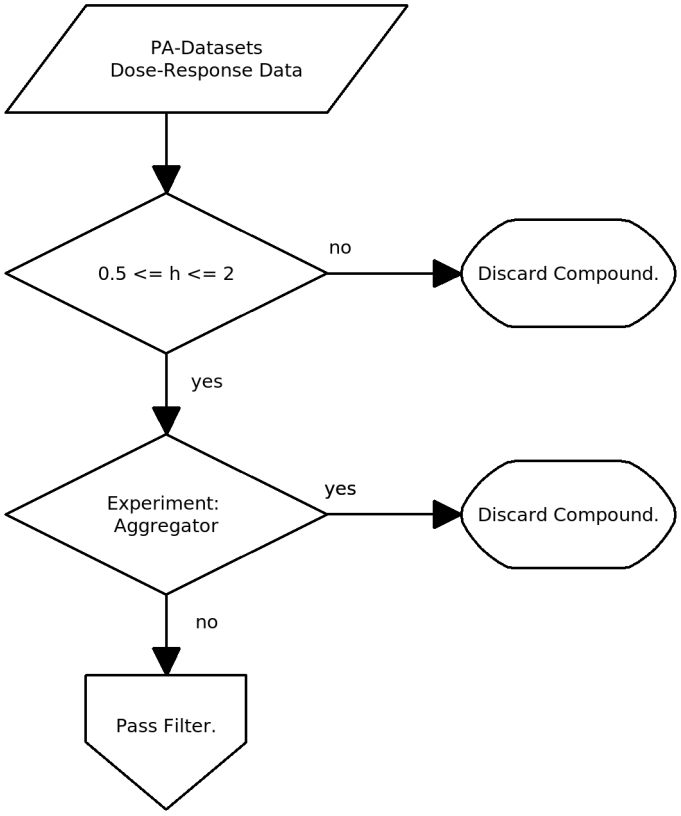
3.2.5.1 Hill Slope Filter

Aggregate formation is often associated with unusual Hill slope values (also referred to as slope factors) in dose-response curves.[9,11,107,111] For competitive inhibition at a single-inhibitor binding site, the Hill slope h is expected to be approximately 1, based on Michaelis-Menten kinetics.[112] Although kinetic theory does not allow a ready application of this expectation to cell based assays or assays screening for allosteric modulation, very large Hill slopes nevertheless raise doubts about the specificity of the observed response. Generally, Hill slopes exceeding 2 are interpreted as harbingers of unspecific activity.[9,111] Hill slopes for the dose response curves of all *PA* compounds were determined. If the Hill slopes were deposited in PCBioAssay, these values were used (Table 3.1). For all other *PA* compounds, Hill slopes were calculated directly from PubChem dose-response data using GraphPad Prism 4.[113] For a *PA* compound to pass the Hill slope filter, its Hill slope h was required to be in the interval $h = [0.5, 2]$. The filter was supplemented by a list of experimentally verified aggregators recently deposited in PCBioAssay (AIDs: 584, 585).[111] Compounds identified as aggregators in this screen were removed from all *PA* datasets. Algorithm 3.1 summarizes the functionality of the Hill Slope Filter.

3.2.5.2 Frequency of Hits (FoH) Filter

In addition to the special case of aggregate formation, bioassays are prone to a range of artifacts caused by unspecific activity of chemical compounds. For cell-based reporter

Algorithm 3.1 Flow-chart of the Hill slope filter. Compounds with undesirable Hill slopes h and compounds that have been experimentally shown to be aggregators are filtered from the *PA* datasets.



gene assays, for instance, this is mainly caused by off-target or cytotoxic effects.[108] These artifacts have been associated with distinct molecular features and cellular actions of compounds showing unspecific activity, which have been termed "frequent hitters".[44] Several studies have tried to identify frequent hitters by first predicting their alleged mode of action as off-target promiscuous binders, aggregators or cytotoxins. Frequent hitters are then flagged based on this prediction.[108,114,115] With a large resource of bioassay data such as PubChem, it is possible to flag a compound as an unspecific binder based on the ratio of the number of assays in which it occurs as a hit and the number of assays in which it was tested. This ratio is termed Frequency of hits (*FoH*). It is expected to be small for specific binders (tested in many assays, but a hit in few assays) and large for unspecific binders (tested in many assays and a hit in most assays). The frequency distribution of *FoH* for a large set of compounds typically features two peaks: one at small values reflecting the population of specific binders and another one at large values of FoH indicating unspecific binders. (Figure 3.2A) In order to distinguish the two populations, the first local minimum of the distribution can be utilized as a conservative, empirical cutoff. At this point, the descent of the FoH distribution after the first peak (specific binders) fades into the ascent to the second peak (unspecific binders). Using piecewise bandwidth optimization, a local polynomial P was fitted to the distribution of *FoH* determined for all PubChem compounds active in at least five bioassays (Figure 3.2B). The first minimum was determined as the second zero-crossing of the first derivative of P at $FoH = 0.26$. (Note, that the first zero-crossing of the derivative occurs at the peak of specific binders.) Consequently, compounds with a *FoH* larger than 0.26 were considered potentially unspecific binders and thus removed from the *PA* datasets.

In this analysis, it is important to take into account, that some assays test for activity against very closely related targets. Whereas for example a compound that was found active in screens against four closely related cholinergic receptors might very well be a real binder to all of them, a compound found active against a kinase, a protease and in two cytotoxicity screens, is highly likely to be unspecific. Thus, it is desirable to correct the *FoH* for the presence of closely related assays in PCBioAssay. As a first measure,

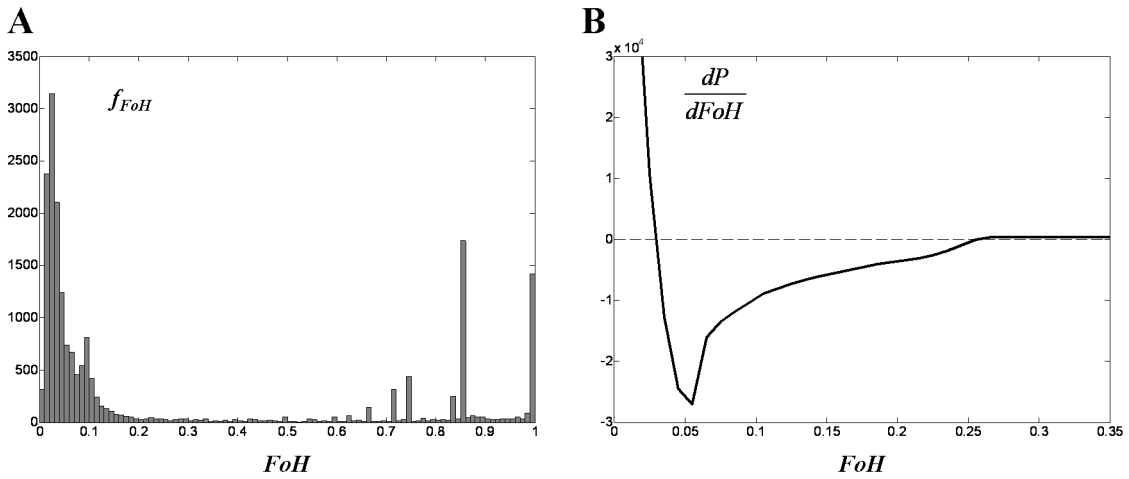


Figure 3.2: (A) Histogram of *FoH* for all compounds active in at least five PCBioAssay HTS screens. The histogram reflects two distinct populations of compounds: The left-hand part of the distribution (small *FoH*) is dominated by specific binders. The right-hand part (large *FoH*) reflects unspecific binders hitting in multiple assays. (B) The first derivative of a polynomial *P* fitted to the histogram has its second zero crossing at $FoH = 0.26$, corresponding to the first minimum of the *FoH* distribution. This was determined as the cutoff beyond which a compound was considered an unspecific binder.

only large scale (>10000 compounds) HTS assays were considered in the FoH analysis, in order to exclude confirmation assays against identical targets from the statistic. The remaining assays were weighted according to the sequence identity of their respective protein targets. Of the 313 HTS assays in PCBioAssay, 163 were associated with protein target information. The respective protein sequences were downloaded from Entrez Protein by a Perl script utilizing Entrez E-Utilities.[16] Using ClustalW[116–118] a multiple sequence alignment was constructed for all sequences. The resulting guide tree was converted into a distance matrix linking all assays by the pairwise percent sequence identity of their respective targets (Figure 3.3). For each compound the set of assays, in which it was active, was determined. Weights were calculated for each assay :

$$w = 1 - \%SI/100; \tag{3.1}$$

with *%SI* the percent sequence identity with the most closely related target associated to one of the assays in the set. All unrelated assays, including assays without protein target annotation, were weighted by 1. Using these weights, a weighted count of assays in which it was found active (*wAAC*) was calculated for each compound. The *FoH* of each

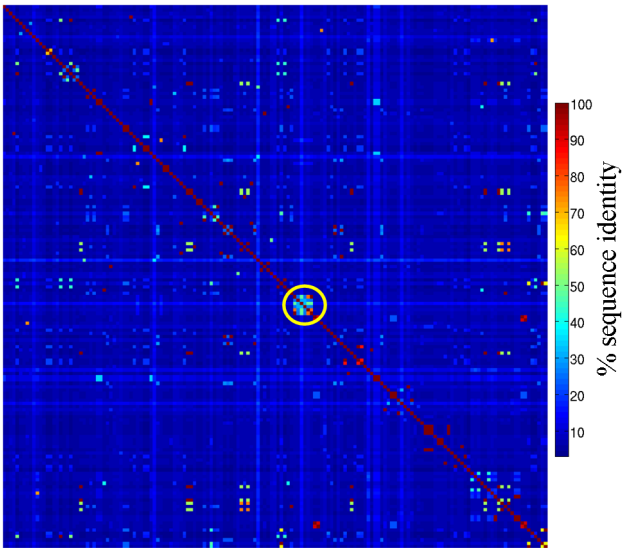


Figure 3.3: Heat-map visualization of the sequence similarity distance matrix of PubChem HTS assays with protein target annotation. Many groups of assays with closely related protein targets are perceptible, most notably a cluster of Ras-related GTPases (highlighted by the yellow circle).

compound was calculated as:

$$FoH = wAAC/TAC; \tag{3.2}$$

with *wAAC* the weighted number of assays in which the compound was found active and *TAC* the number of assays in which it was tested. This weighted *FoH* score was actually used throughout the analysis of the PubChem datasets.

3.2.5.3 Autofluorescence and Luciferase Inhibition Filter

Assays based on optical detection are often affected by the chromo/fluorogenic properties of some compounds. Recently, a large scale fluorescence spectroscopic profiling of PubChem substances has been carried out and been reported in PCBioAssay.[110] Compounds that were found to exhibit undesirable properties in this profiling (AIDs: 587, 588, 590, 591, 592, 593, 594) were removed from the PA datasets. In a similar fashion, a recent screen tested a large set of PubChem substances for their potential to inhibit Luciferase.[109] Compounds found active in this screen (AID 411) were also removed from the PA datasets.

3.2.6 Potential False Negatives in the Datasets of Decoys

In addition to false positives, which are addressed by the assay artifacts filter, HTS experiments are also affected by false negatives, i.e. active compounds are falsely designated as inactive. This constitutes a potential source of error for the decoy sets presented here. In contrast to false positives, the data in PCBioAssay provides no means to detect potential false negatives. Furthermore, it is not possible to apply statistical methods such as bootstrapping in order to get an estimate of the error introduced by false negatives, since there is no way to reasonably estimate the rate of false negatives in PubChem single dose HTS assays. In the future, results of quantitative high throughput screening (qHTS)[8] experiments may provide numerical data that might form the basis of such calculations. At the moment however, these results are specific for the respective target, detection method, compound library and tested compound concentration. Thus, they cannot be applied to the decoy sets utilized here. (Personal communication: *Christopher P. Austin*, Director NIH Chemical Genomics Center, National Institutes of Health and *Douglas Auld*, Group Leader, Genomic Assay Technologies, NIH Chemical Genomics Center, National Institutes of Health) In order to get at least some idea about the validity of the selected decoys, a similarity search employing simple descriptors and MAX-rule data fusion was performed on each MUV decoy set, using the complete set of actives as query. For each MUV dataset, the five decoys most similar to the actives were recorded. For each of the resulting 85 compounds, an extensive literature research was performed using SciFinder Scholar.[119] The results are summarized in Tables B.1 and B.2 (Appendix). In summary, no reports could be found in the literature that suggested a specific activity of one of the decoys against the target of its respective MUV dataset. Although this evidence is at best anecdotal, together with the experimental result of inactivity in the HTS assay, it justifies some confidence about the inactivity of MUV decoys. Especially compared to traditional VS benchmark datasets, in which the inactivity of the decoys is merely assumed without any experimental evidence whatsoever, this constitutes a considerable improvement.

3.2.7 Chemical Space Embedding Filter

In order to prevent artificial enrichment, decoys are selected to be similar to the set of actives regarding "simple" molecular properties. Usually, this is achieved by selecting a set of neighbors for each active **a** from a set of potential decoys (Figure 3.4A). However, if chemical space around **a** is devoid of decoys, no selection of decoys is possible, that can prevent artificial enrichment (Figure 3.4B). Actives must be well embedded in decoys to allow unbiased decoy set design. Thus, actives inadequately embedded in decoys were removed from the *PA* datasets by a "chemical space embedding filter". In order to quantitatively define "good embedding", a comprehensive sample of drug-like chemical space was compiled. Compounds were pooled from DrugBank,[120] Prous Drugs of the Future,[121] the Sigma-Aldrich chemistry catalog[122] and the MDDR.[69] This collection, which comprised 372021 unique compounds, will be referred to as the "chemical space sample" in the remainder of the text. With exception of the MDDR, the chemical space sample was downloaded in SD-Format[80] from PubChem. Extremely large compounds (i.e. proteins) were filtered from the sample using MOE sdwash.[34] Small fragments and counter ions were removed and 3-dimensional structures generated using CORINA.[81] It is safe to assume that all actives are well embedded in this comprehensive collection of compounds. All datasets and the chemical space sample were encoded by simple descriptors. For each *PA* dataset, a random sub-sample of the same size as the corresponding *PD* dataset was drawn from the chemical space sample. For each active **a**, the distance to the 500^{th} nearest neighbor in the random sample was determined. This was repeated 100 times and the 90^{th} percentile d_{90} was recorded as the 90% confidence boundary for a good embedding of **a**. (Figure 3.4C) The distance to the 500^{th} nearest neighbor in the decoy set, d_{Decoys}, was determined in an analogous fashion. Actives were considered inadequately embedded in decoys and thus removed from the *PA* datasets, if d_{Decoys} was larger than d_{90}. (Figure 3.4D) The distance to the 500_{th} nearest neighbor in the decoy set was chosen as a criterion for chemical space embedding, because 500 decoys were selected for each active in the process of MUV dataset design. (actives: $k = 30$, decoys: $d = 15000$)

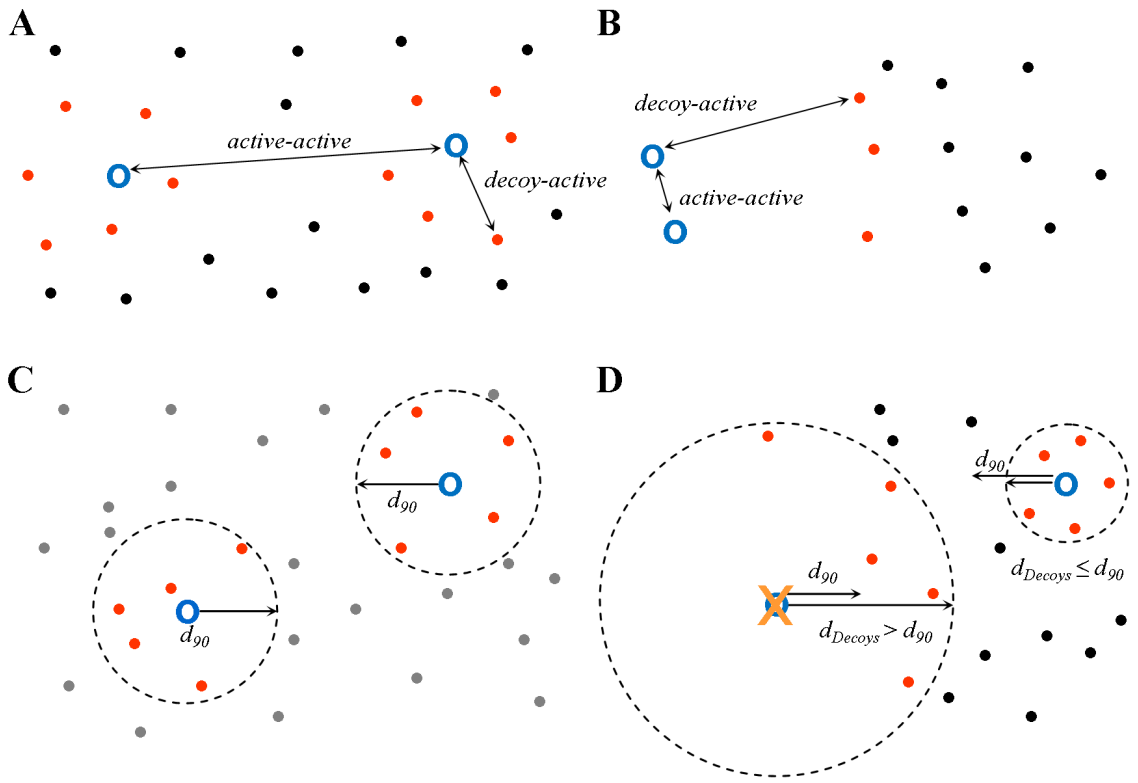


Figure 3.4: (A) A set of similar decoys (red) is selected for each active. Since active-active distances are generally larger than decoy-active distances, artificial enrichment is prevented. (B) If actives are inadequately embedded in decoys, decoy-active distances result, that are larger than active-active distances even for the most similar decoys. Artificial enrichment is the consequence. (C) For each active (blue circles) the distance to the 500^{th} nearest neighbor in a representative sample of compounds (grey dots) is determined. This corresponds to the radius of a hypersphere incorporating the 500 nearest neighbor compounds (red dots, five nearest neighbors are used here for the sake of clarity). d_{90} is determined as a 90% confidence boundary of this distance in 100 samples of compounds. (D) If an active is located in a region of chemical space that is significantly devoid of decoys (black dots), the hypersphere containing the 500^{th} nearest decoys, d_{decoys} has a radius decoys larger than d_{90}. Such compounds are inadequately embedded in decoys and is excluded from the *PA* dataset.

3.2.8 Descriptors

Simple descriptors (see Section 2.2.3) were calculated for dataset analysis and design. In order to validate how other descriptors perform on the datasets, they were encoded by three additional classes of descriptors: SESP a class of versatile, alignment-independent 2D topological indices based on atom pairs,[38] MOE molecular properties descriptors[34] and MACCS structural keys.[123] Bias introduced by the 3D conformation generator CORINA was excluded by using only the 2D class of MOE descriptors. Since the numerical values of properties in descriptors have significantly different ranges, all *PA/PD* descriptor matrices, except the ones encoded by MACCS keys, were autoscaled column-wise by subtraction of the mean and division by the standard deviation of the respective column in the chemical space sample. Columns with a standard deviation of 0 were removed from all descriptor matrices including MACCS keys. After this pretreatment, the descriptor matrices of simple, MACCS, MOE and SESP had a dimensionality of 17, 154, 184 and 418, respectively. In order to reduce noise, principal components analysis (PCA)[82] was applied to the descriptor matrices of MOE and SESP. The loadings (eigenvectors) were derived from a singular value decomposition (SVD)[124] of the chemical space sample (see Section 3.2.7) encoded by the respective descriptors. An analysis of the resulting eigenvalues showed that $> 90\%$ of the total variance could be explained by the first 70 components for MOE and the first 94 components for SESP, respectively. Thus, the first 70 (MOE) and the first 94 (SESP) scores from the PCA were used as the final descriptors.

3.2.9 Spatial Statistics for Benchmark Dataset Design

3.2.9.1 Preliminary Experiments for the Determination of t_i

For the analysis of chemical datasets by Refined Nearest Neighbor methods it is important to take account of the fact that the distribution of nearest neighbor distances in a dataset greatly depends of the respective descriptor space. Depending on the dimensionality of a descriptor (e.g. MOE $p = 70$ vs. Simple $p = 17$, p: number of components) and the nature of the descriptor values (e.g. MOE $[-\infty;\infty]$ vs. MACCS $\{0,1\}$), distances to

the nearest neighbor compound might be generally larger in one descriptor space than in another. Figure 3.5A shows graphs of $G(t)$ for 100 random sub-samples taken from the chemical space sample based on MOE and autoscaled simple descriptors, respectively. Apparently nearest neighbor distances are considerably larger in MOE descriptor space than in simple descriptor space for the same sub-samples of compounds. As long as the topology of two or more datasets is assessed based on a single descriptor representation, this effect is without consequence. The effect becomes critical, however, if the topology of a dataset is compared across different descriptor spaces. Figure 3.5B shows graphs of $G(t)$ for the PubChem bioactivity dataset AID 846 in simple and MOE descriptor space respectively. At first sight, it seems that the dataset is considerably more clumpy in simple descriptor space. However, when comparing the graphs to the ones shown in Figure 3.5A, it is evident that this is mostly caused by the generally shorter nearest neighbor distances in simple descriptor space.

Especially when assessing a descriptor's mapping performance (see Section 2.3.5), it is essential to assess a descriptor's ability to convert the specific molecular features of a dataset of actives into. Clumpiness inherent to the descriptor space would distort the results of such examinations. This can be prevented by scaling the range of distances t_i, which is the basis for the calculation of $G(t)$ and $F(t)$, by a factor that is proportional to the expansion of the respective descriptor space. In order to do so, a maximum distance t_{max} must be determined as:

$$t_{max} = c * f_{spatial}; \tag{3.3}$$

where $f_{spatial}$ constitutes the descriptor space expansion factor. c is a constant that can be chosen arbitrarily, with the sole condition of being large enough to accommodate all nearest neighbor distances in the datasets under examination. Furthermore, a resolution Δt must be determined so that t_i is incremented in an equal number of fractions of t_{max} for all descriptor spaces:

$$\Delta t = \frac{1}{r} * t_{max}; \tag{3.4}$$

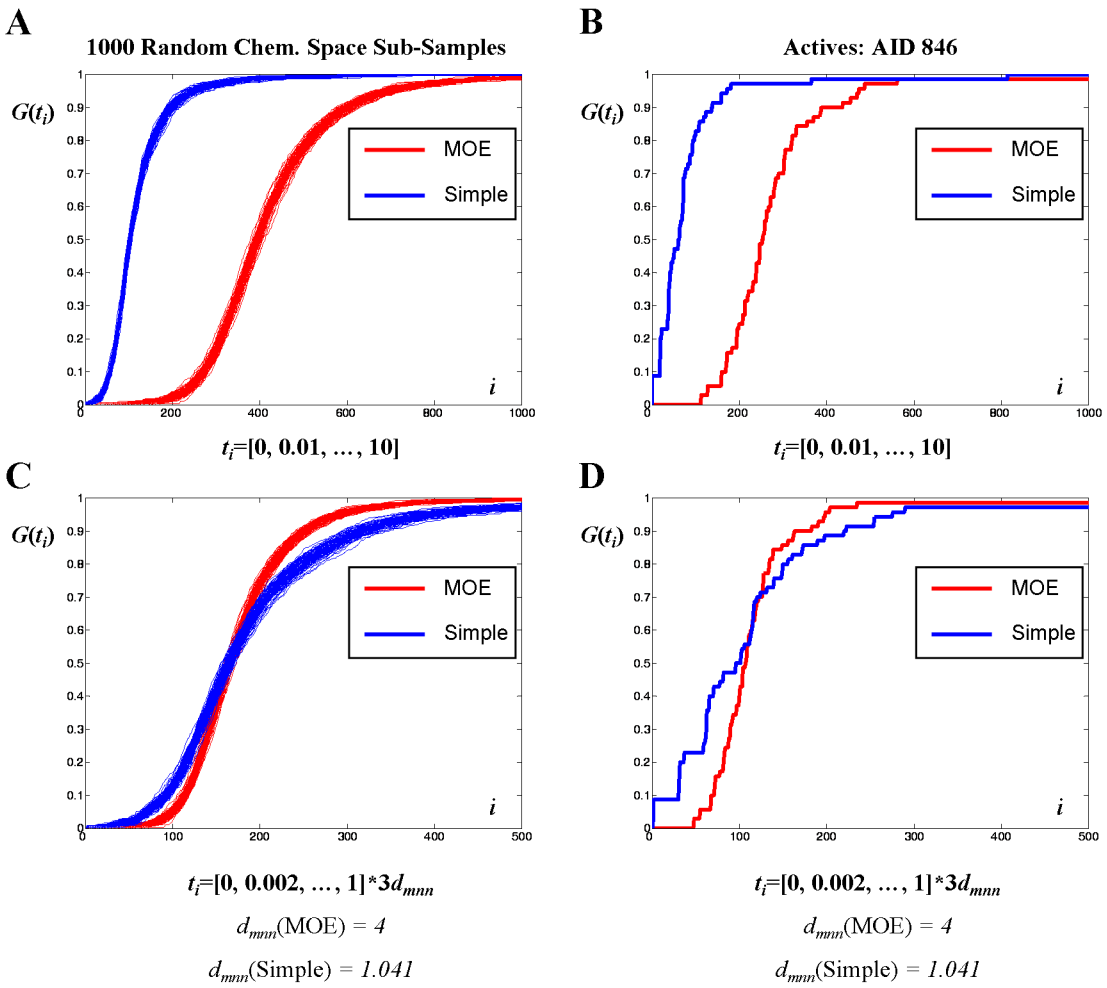


Figure 3.5: (A) Graphs of $G(t)$ for 100 random chemical space sub-samples encoded by Simple and MOE descriptors, respectively. Nearest neighbor distances are inherently larger in MOE descriptor space than in Simple descriptor space. (B) Analyzing the different levels of clumping of the representations of dataset AID 846, it is difficult to distinguish mapping performance from effects caused by differences in descriptor space expansion. (C) Scaling the range of distances t_i using the descriptor space expansion factor d_{mnn}, curves of $G(t)$ are generated that are comparable across descriptor spaces. (D) When applied to the curves of $G(t)$ for dataset AID 846 in both descriptor spaces, scaling of t_i reveals that the nearest neighbor distances occurring in this datasets are roughly equal in both descriptor spaces.

The resolution factor $\frac{1}{r}$ can be chosen arbitrarily based on computation time considerations, as long as it is sufficiently small to capture the differences in nearest neighbor distances in the examined datasets. The range of distances t_i is then given as:

$$t_i = [0, \Delta t, ..., t_{max}]; \tag{3.5}$$

In order to determine a reasonable choice for the descriptor space expansion factor $f_{spatial}$, preliminary experiments were carried out. Based on the chemical space sample, several measures describing a descriptor space's expansion were calculated for each descriptor space. These included the maximum distance of two compounds in the sample d_{max}, the mean of all pairwise distances d_{Mpw}, the median of all pairwise distances d_{mpw}, the mean nearest neighbor distance d_{Mnn} and the median nearest neighbor distance d_{mnn}. 1000 random sub-samples were extracted from the chemical space sample. $G(t)$was calculated for all sub-samples with

$$t_{max} = 3 * f_{spatial};$$

$$f_{spatial} = \left\{ d_{max}, d_{Mpw}, d_{mpw}, d_{Mnn}, d_{mnn} \right\};$$

$$\Delta t = \frac{1}{500} * t_{max};$$

for all descriptors used in this Chapter. (Simple, MOE, SESP, MACCS) By analyzing the resulting curves visually, the mean nearest neighbor distance d_{Mnn} and the median nearest neighbor distance d_{mnn} of the chemical space sample in the respective descriptor spaces were determined as the factors best suited for the generation of comparable curves across all descriptor spaces. d_{Mnn} and d_{mnn} were found to be identical up to the fourth decimal. The median nearest neighbor distance d_{mnn} was chosen as the descriptor space expansion factor for the experiments described in the following section, because it is robust and does not depend on extreme observations. Figure 3.5C shows the

graphs of $G(t)$ for the same 100 random sub-samples as mentioned above calculated with $t_i = [0, \frac{1}{500}, ..., 1] * 3 * d_{mnn}$ for Simple and MOE descriptors. The distributions of nearest neighbor distances in both descriptor spaces are now comparable. Consequently the comparison of the mappings of dataset AID 846 in the two descriptor spaces is much more informative. (Figure 3.5D)

3.2.9.2 Spatial Statistics for PubChem Datasets

Based on the findings described above, t_{max} and Δt were set to:

$$t_{max} = 3 * d_{mnn,Simple}; \tag{3.6}$$

$$\Delta t = \frac{1}{500} * t_{max}; \tag{3.7}$$

for the PubChem bioactivity datasets. The values of $c = 3$ and $r = \frac{1}{500}$ were found sufficient for the topologies encountered throughout this Chapter. Using the chemical space sample (see Section 3.2.7), $d_{mnn,Simple} = 1.041$ was determined as the median of all nearest neighbor distances in simple descriptor space. Thus, $G(t)$, $F(t)$ and $S(t)$ and the respective numerical integrals ΣG, ΣF and ΣS were calculated according to Eqs. 2.4, 2.5, 2.6, 2.7, 2.8 and 2.9 utilizing Algorithms 2.1 and 2.2 with $t_i = [0, 0.006, 0.012, ..., 3] * 1.041$.

3.2.10 Design of MUV Datasets

The goal of MUV design is to generate sets with a spatially random distribution of actives and decoys in simple descriptor space. This was accomplished by a two step procedure. First the datasets of actives were adjusted to a common level of spread. Subsequently, decoy sets were selected with a common level of separation from the actives. Subsets of $k = 30$ actives with the maximum spread possible in each *PA* dataset were generated using the well established Kennard-Stone[125] algorithm. Since this algorithm generates the maximum spread for each individual dataset, the maximum common level of spread

to which all datasets can be adjusted, is the lowest level of spread observed among all Kennard-Stone subsets. This corresponds to the maximum value of ΣG among the subsets of $g = 312$. A row-exchange algorithm[91] (Algorithm 3.2, Source code: Appendix D.2.3) was applied to reduce the spread of all datasets with $\Sigma G > 312$ to a level of $\Sigma G \approx 312$. With the respective Kennard-Stone subsets of actives as a starting design, compounds were exchanged with the remaining *PA* compounds until the objective function

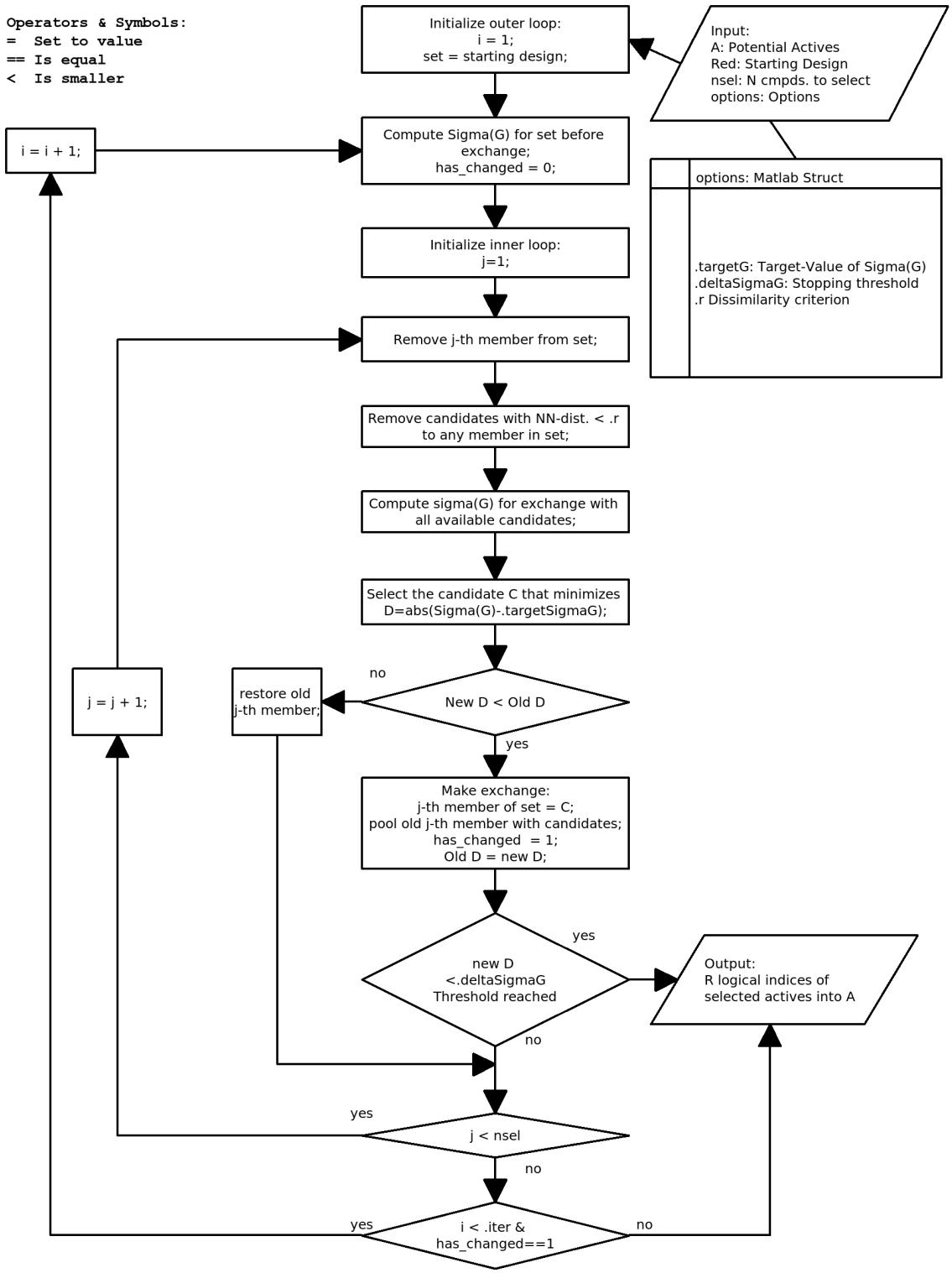
$$D = \mid g - \Sigma G_{Dataset} \mid; \tag{3.8}$$

reached values of $D \leq 2$, with $\Sigma G_{Dataset}$ representing ΣG of the dataset at the respective iteration, or if no exchange could be made that would further enhance $\Sigma G_{Dataset}$. As an additional constraint, only compounds were allowed to be selected that had a nearest neighbor distance larger than $r = 0.8$ to the compounds already in the selection. The dissimilarity constraint r is used in analogy to the well-established OptiSim algorithm[126] for diverse subset selection.

Both the cutoff value of $D \leq 2$ and $r = 0.8$ were determined empirically by preliminary experiments. While the main reason for the choice of $D \leq 2$ was keeping computing time at a reasonable level, the value of $r = 0.8$ is critical for the properties of the resulting datasets. Here, larger values of r would constitute an excessively harsh dissimilarity constraint, i.e. such values would exclude too many *PA* compounds from selection for the datasets and consequently render the selection of 30 actives impossible. Smaller values of r on the other hand would allow the selection of compounds that are very similar to compounds already in the set. Since ΣG captures no information about the shape of the curve of $G(t)$, the design criterion of $\Sigma G = 312$ can also be fulfilled if $G(t)$ increases early but with a flat slope, i.e. by a dataset that contains a small number of very similar compounds. Such datasets however would be subject to considerable analogue bias. Therefore r should be set to the maximum value at which the necessary number of actives (here $k = 30$) can still be selected for all datasets. For the PubChem bioactivity datasets used here, this value was $r = 0.8$.

For actives and decoys to exhibit a spatially random distribution, ΣF must be equal

Algorithm 3.2 Row-Exchange Algorithm for the Selection of *nsel* Actives with Given Σ*G* from a Larger Set of Potential Actives A.



to ΣG. Therefore, a starting design of decoys was generated by selecting the 500 most similar decoys for each active, resulting in a set of $d = 15000$ decoys for each dataset. ΣF was adjusted to $f = g = 312$ by a genetic algorithm (Algorithm 3.3, Source code: D.2.4) with

$$D = \mid f - \Sigma F_{Dataset} \mid; \tag{3.9}$$

as the fitness function, with $\Sigma F_{Dataset}$ representing ΣF of the dataset at the respective iteration. Convergence was reached if $D \leq 2$ or D remained constant for more than 10 iterations. Again, $D \leq 2$ was determined as a reasonable fitness cutoff by preliminary experiments.

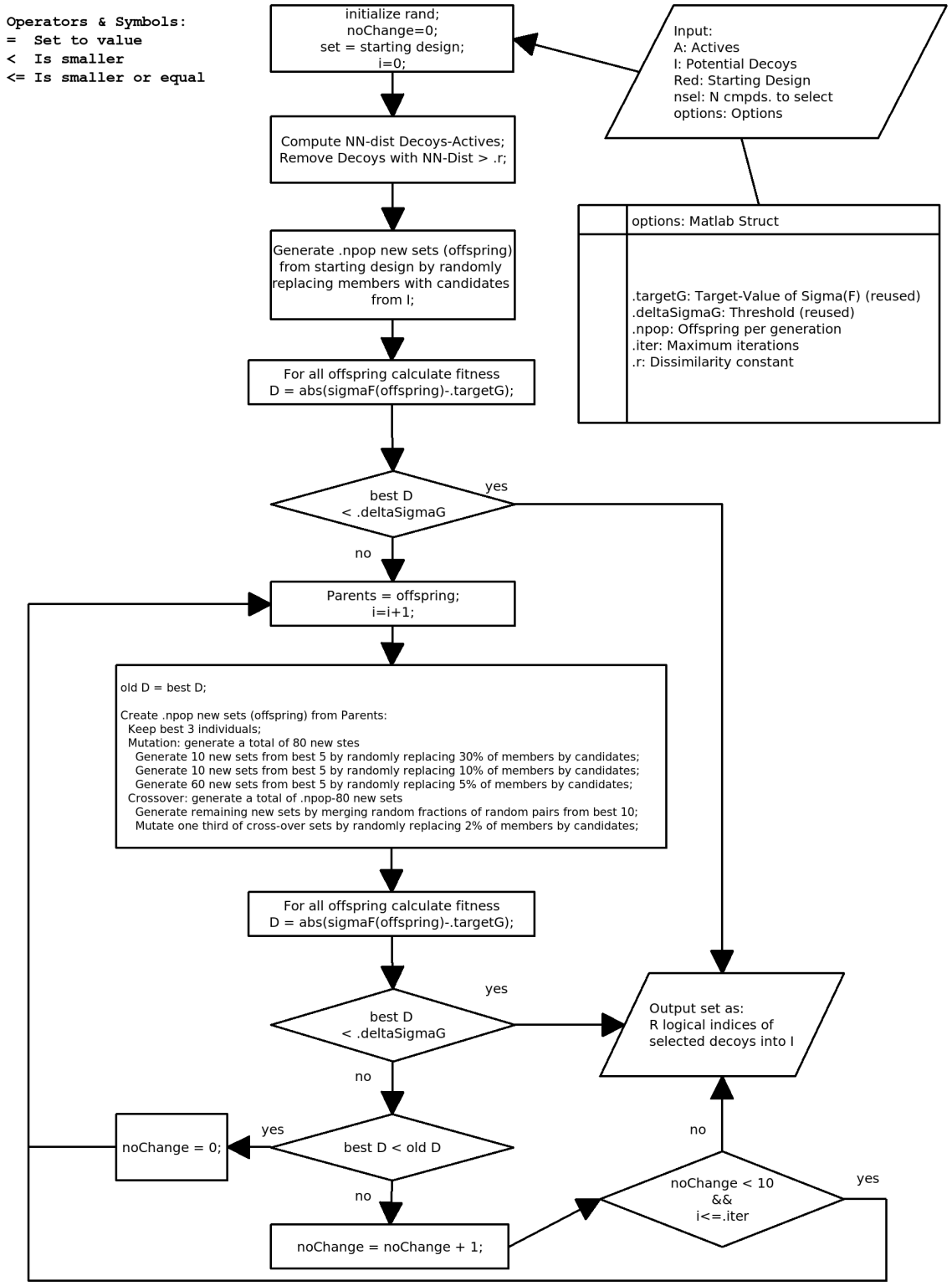
3.2.11 Retrospective virtual screening simulations

In order to demonstrate the utility of the MUV datasets in VS validation experiments, retrospective VS simulations were carried out. The datasets of actives and decoys were encoded by all descriptors (simple, MOE, SESP, MACCS). For each run, a query of one or ten compounds, respectively, was chosen randomly from the datasets of actives. The remaining actives were pooled with the respective decoys to form the validation set. For each MUV dataset of actives, 100 such random splits were generated, in order to obtain a mean value of VS performance that is not affected by the random choice of the query molecules. Similarity was measured by the Euclidean distance and the validation sets were ranked accordingly. For queries consisting of ten compounds, MAX-rule data fusion[47,48] was applied to the ranking.

3.2.12 Figures of Merit (FoM) for Virtual Screening Performance

VS performance was measured by the area under the receiver operating characteristic curve (AUC_{ROC}). Additionally, the ability for early recognition of active compounds was quantified by the fraction of retrieved actives (Retrieval Rate, $RTR_{1\%}$) in the first percent of the ranked validation set. The mean areas under the receiver operating characteris-

Algorithm 3.3 Genetic Algorithm for the Selection of *nsel* Decoys with Given ΣF Based on a Set of Actives A and a Set of Potential Decoys I.



tic curves and mean Retrieval Rates obtained from the 100 random query / validation set splits generated for each dataset, will be denoted $mean(AUC_{ROC})$ and $mean(RTR_{1\%})$ throughout the text.

3.2.13 Unique Molecular Frameworks

A relatively recent concept regarding the validation of VS methods is the notion that good virtual screening algorithms should be able to perform a so-called "*scaffold hop*".[37,127] Given a certain query molecule, a method should be able to extract actives from a database that feature a novel scaffold, i.e. a chemical backbone different from that of the query compound. In order to test for a method's ability to perform scaffold hops, a benchmark dataset of actives is required to contain a large number of distinct scaffolds, which is equivalent to the requirement that each scaffold class in the dataset is represented by only a small number compounds. More specifically, the ratio of compounds per scaffold class should not considerably exceed 1.

Although "scaffold hopping" is one of the central concepts of modern virtual screening research, many definitions of the term "scaffold" exist in the medicinal chemistry literature.[37,127–130] Of these, the definition of scaffolds as "molecular frameworks" (Figure 3.6) proposed by Bemis and Murcko in their seminal paper[128], is most widely adopted. In their algorithm, a 2-dimensional representation of a molecule's structure is first converted to a molecular graph by ignoring all atom and bond types. In the second step all side chain atoms and bonds are removed to produce the molecular framework.

In order to determine the number of unique molecular frameworks in the PubChem bioactivity datasets before and after MUV design, all compounds in a dataset were converted to molecular frameworks using an in-house MOE SVL script.[34] (Source code: Enclosed CD-ROM) Compounds were grouped according to their molecular framework and the number of unique groups was recorded for each dataset.

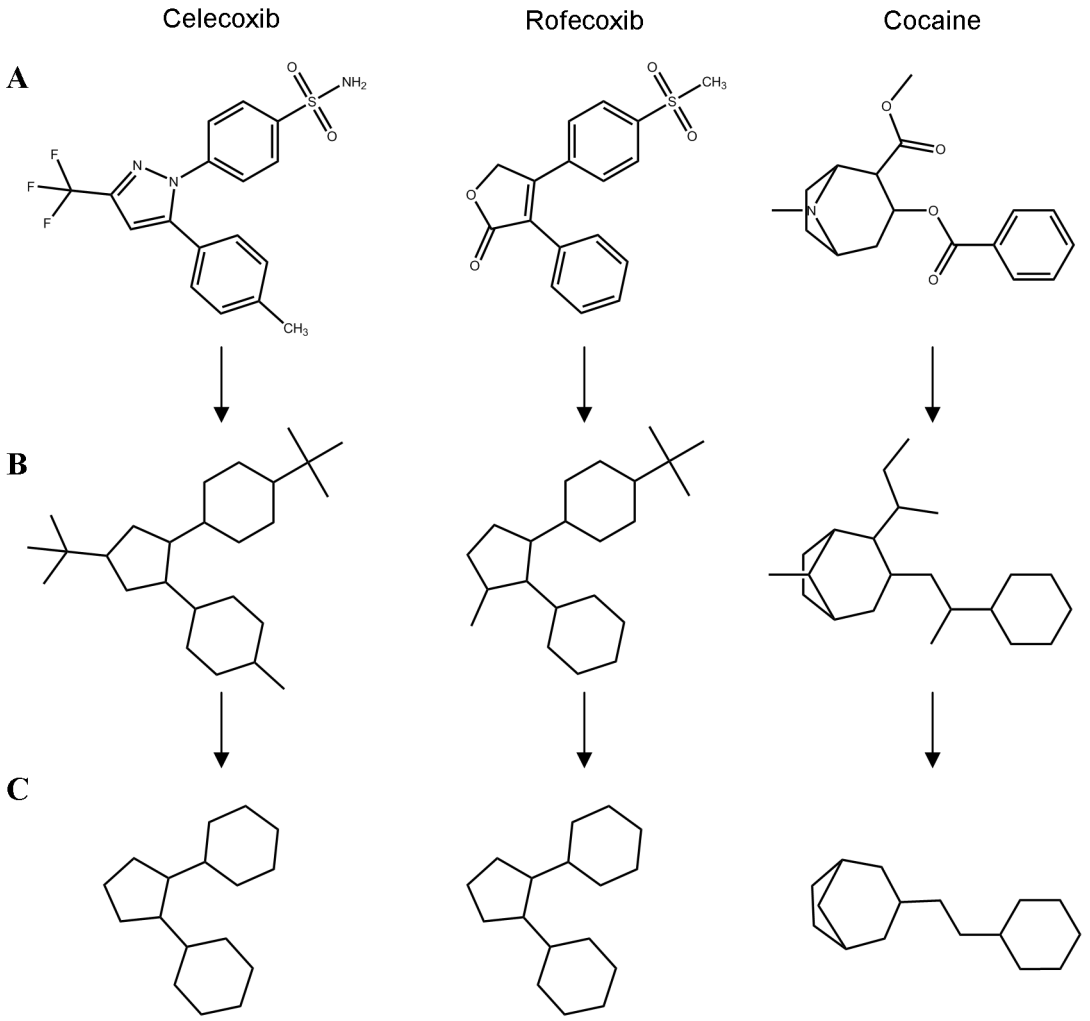


Figure 3.6: Based on the 2-dimensional representation of a molecule's structure (A) a molecular graph is generated by ignoring all atom or bond types. (B) A molecular framework is generated by "pruning" the molecular graph for all side chain atoms and bonds. (C) After this treatment, both Celecoxib and Rofecoxib are identified as members of the Coxib scaffold class, which is distinct from the molecular framework of Cocaine. Consequently, a dataset consisting of these three compounds would contain two unique molecular frameworks and thus feature an average ratio of 1.5 compounds per scaffold class.

3.3 Results and Discussion

3.3.1 Bioactivity Datasets Extracted from PubChem

From the bioactivity data available in PCBioAssay, 18 pairs of assays against pharmaceutically relevant targets were selected. Each of these pairs included a primary high-throughput screen and a low-throughput confirmation assay against the same target. Active compounds from the confirmation assays were used as the dataset of potential actives (*PA*) for the subsequent design steps. In a complementary fashion, inactive compounds from the corresponding primary screens were used as datasets of potential decoys (*PD*). An overview of the respective pairs of bioassays is shown in Table 3.1. (Section 3.2.4) High standards regarding the specificity of the bioactivities in the datasets of actives were enforced by selecting only low-throughput confirmatory assays with associated dose-response information and EC_{50} values as *PA* datasets. An assay artifacts filter further removed compounds with a potential for unspecific activity, including aggregators, promiscuous binders and compounds interfering with optical detection methods (Table 3.3). In addition to the datasets of actives, the specificity of the datasets of decoys was enforced by utilizing only compounds for potential decoy (*PD*) datasets, whose inactivity against the respective target was experimentally determined by HTS. Although decoys that were false negatives in the HTS cannot be detected by these procedures, the level of confidence in the inactivity of the MUV decoys is still higher than for benchmark datasets that merely assume compounds without any annotated activity to be inactive.

3.3.2 MUV Benchmark Datasets: General Properties

Subsets of $k = 30$ actives and $d = 15000$ decoys were selected from each PA/PD pair of datasets, that were as close to spatial randomness as possible. (see Sections 3.2.10 and 3.3.3) The resulting datasets contain a remarkably high number of distinct molecular scaffolds (Table 3.2). On average, MUV datasets contain only 1.16 compounds per scaffold class, a ratio that effectively eliminates analogue bias. Here, scaffolds were defined as reduced molecular graphs as proposed by Bemis and Murcko.[128] Further, MUV datasets

Table 3.2: Effect of MUV Design Strategy on PA Datasets.

		Number of Compounds in Dataset				Unique Molecular Frameworks	
Target	AID	*PA*	Assay Artifacts Filter[a]	Chem. Space Emb. Filter[a]	MUV Design[a,b]	*PA*	MUV
S1P1 rec.	466	223	185 (-38)	180 (-5)	30	127	28
PKA	548	62	51 (-11)	50 (-1)	30	37	27
SF1	600	213	71 (-142)	70 (-1)	30	47	24
Rho-Kinase2	644	67	58 (-9)	57 (-1)	30	39	27
HIV RT-RNase	652	370	178 (-192)	169 (-9)	30	121	27
Eph rec. A4	689	80	58 (-22)	56 (-2)	30	48	29
SF1	692	75	37 (-38)	37 (-0)	30	36	30
HSP 90	712	90	46 (-44)	44 (-2)	30	32	27
ER-a-Coact. Bind. Inh.	713	221	98 (-123)	92 (-6)	30	81	26
ER-ß-Coact. Bind. Inh.	733	194	104 (-90)	101 (-3)	30	78	28
ER-a-Coact. Bind. Pot.	737	64	48 (-16)	42 (-6)	30	39	28
FAK	810	110	71 (-39)	62 (-9)	30	51	28
Cathepsin G	832	65	51 (-14)	51 (-0)	30	31	24
FXIa	846	70	61 (-9)	60 (-1)	30	31	21
S1P2 rec.	851	54	28 (-36)	23 (-5)	23	16	16
FXIIa	852	99	81 (-18)	80 (-1)	30	39	24
D1 Rec.	858	226	140 (-86)	138 (-2)	30	106	24
M1 Rec.	859	234	149 (-85)	133 (-16)	30	103	29

[a)] Number of compounds left in dataset after application of the respective filter/design step.

[b)] $k = 30$ actives was the design criterion for MUV datasets.

Table 3.3: Average Number of Lipinski's Rule-of-5 Violations per Compound in MUV Datasets

Target	AID	Actives	Decoys
S1P1 rec.	466	0.07	0.02
PKA	548	0	0.01
SF1	600	0.03	0.01
Rho-Kinase2	644	0	0.01
HIV RT-RNase	652	0.10	0.01
Eph rec. A4	689	0.23	0.06
SF1	692	0.07	0.01
HSP 90	712	0	0
ER-a-Coact. Bind. Inh.	713	0.03	0.02
ER-ß-Coact. Bind. Inh.	733	0.03	0.02
ER-a-Coact. Bind. Pot.	737	0.07	0.03
FAK	810	0.10	0.04
Cathepsin G	832	0	0.01
FXIa	846	0	0.05
S1P2 rec.	851	0.13	0.08
FXIIa	852	0	0.05
D1 rec.	858	0	0.01
M1 rec.	859	0	0.01

provide a good representation of drug-like chemical space. As shown by Table 3.3, violations of Lipinski's Rule of 5 occur with very low frequency, both in the datasets of actives and decoys. Only 23 compounds in the *PA* dataset of S1P2 receptor inhibitors (AID 851) passed all filters, a number that is insufficient for MUV design of the dataset of actives. The dataset is kept here for illustrative purposes but will not be part of the final MUV collection, which consequently consists of 17 benchmark datasets.

3.3.3 Spatial Statistics Analysis of MUV Datasets

Table 3.4 and Figure 3.7 show an overview of dataset clumping in simple descriptor space before and after MUV dataset design. Before MUV design, all PA/PD datasets show considerable clumping indicated by negative values of ΣS. Over-optimistic validation results would be the consequence, if these datasets were used for VS validation without further design. MUV datasets on the other hand, exhibit ΣS values close to 0, indicating mildly dispersed distributions of actives and decoys close to spatial randomness. With only small

Table 3.4: Measures of Dataset Topology for PubChem Bioactivity Datasets

		No Design			MUV		
		ΣS	ΣG	ΣF	ΣS	ΣG	ΣF
S1P1 rec.	466	-46.12	368.53	322.41	2.08	311.77	313.84
PKA	548	-114.64	374.92	260.28	1.22	311.67	312.88
SF1	600	-39.53	337.86	298.33	1.11	310.67	311.78
Rho-Kinase2	644	-95.66	375.88	280.22	1.18	311.80	312.98
HIV RT-RNase	652	-64.05	369.64	305.60	0.33	310.60	310.93
Eph rec. A4	689	-61.49	326.45	264.96	0.51	311.50	312.01
SF1	692	-51.48	333.32	281.84	12.24	300.77	313.01
HSP 90	712	-65.05	330.98	265.93	10.90	308.40	319.30
ER-a-Coact. Bind. Inh.	713	-9.80	317.92	308.13	1.39	312.00	313.39
ER-ß-Coact. Bind. Inh.	733	-15.91	322.61	306.70	2.91	312.00	314.91
ER-a-Coact. Bind. Pot.	737	-43.05	318.12	275.07	0.92	312.07	312.99
FAK	810	-48.14	321.66	273.52	-1.83	311.07	309.23
Cathepsin G	832	-103.71	383.04	279.33	-3.58	315.83	312.25
FXIa	846	-138.68	397.72	259.04	2.07	311.10	313.17
S1P2 rec.	851	-70.14	277.04	206.90	25.87	266.77	292.64
FXIIa	852	-118.86	393.99	275.13	-1.33	311.77	310.44
D1 Rec.	858	-14.75	339.14	324.39	1.62	311.30	312.92
M1 Rec.	859	-38.62	354.87	316.25	1.46	311.50	312.96

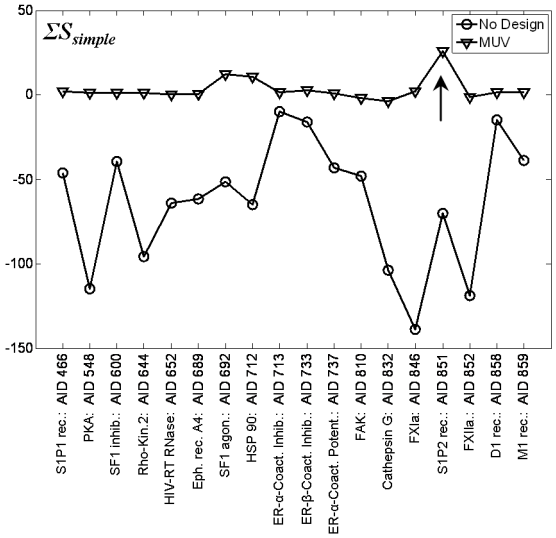


Figure 3.7: Effect of MUV design on dataset topologies in simple descriptor space. Without MUV design all datasets show clumping (negative values of ΣS_{simple}), indicating the potential for benchmark dataset bias. MUV datasets exhibit ΣS_{simple} values close to 0, indicating spatial randomness of actives and decoys. Dataset AID 851 (arrow) does not fulfill the criteria. No design is feasible on the respective dataset of actives, since it contains less than 30 compounds after application of the assay artifacts filter.

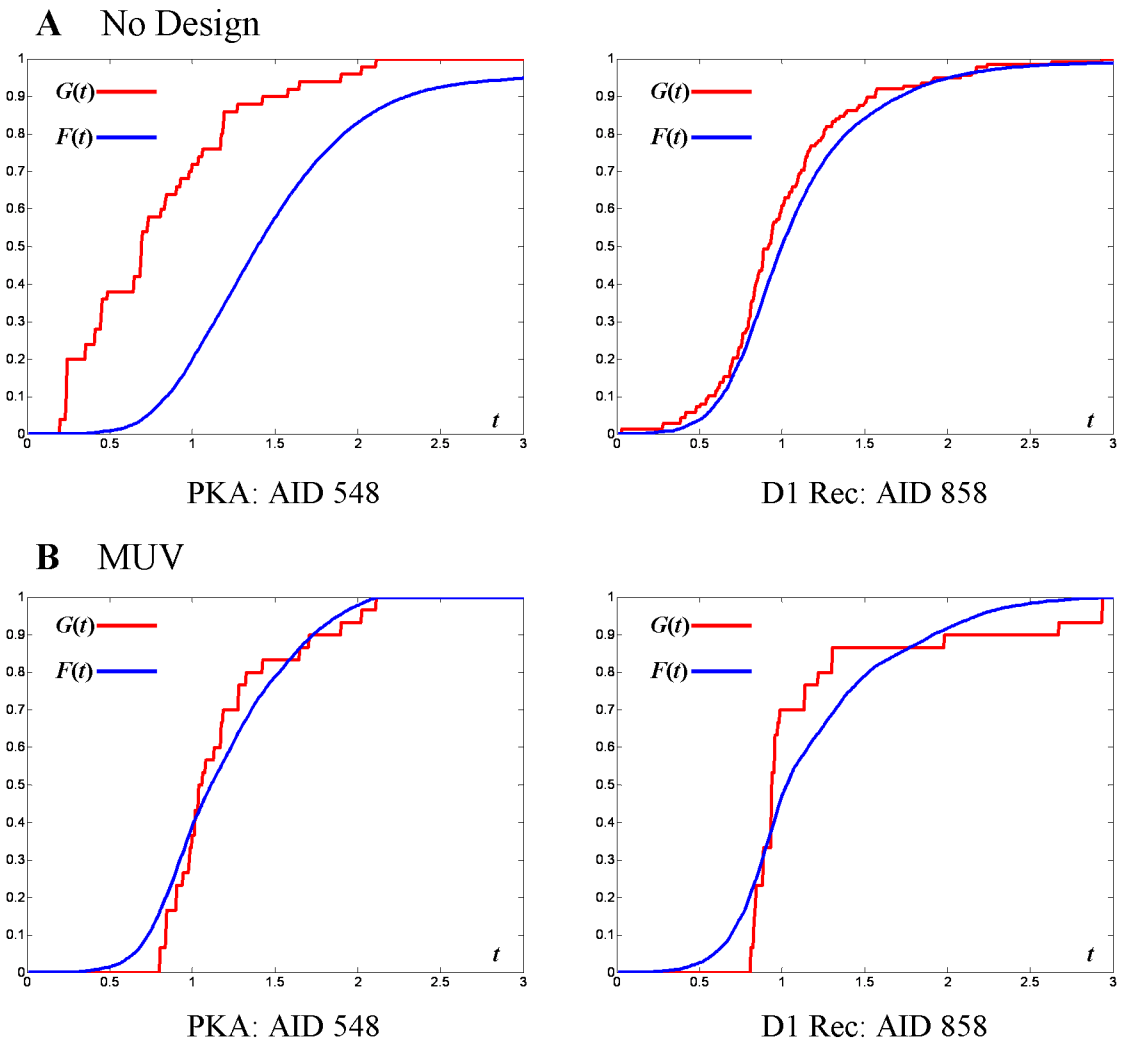


Figure 3.8: Effect of MUV design on the topology of benchmark datasets AID 548 and AID 858 in simple descriptor space. (A) Different degrees of dataset clumping are observable in the raw datasets. The large shift of $F(t)$ relative to $G(t)$ in AID 548 indicates extensive clumping. (B) MUV design reduces the differences in topology between the datasets. Both, $G(t)$ and $F(t)$, show similar curves for both datasets. Comparing $G(t)$ to $F(t)$, it is obvious, that $G(t)$ is coarsely discretized. This is caused by the much smaller number of active compounds used for the calculation of $G(t)$ compared to the very large number of decoys, which are the basis for the calculation of $F(t)$.

variations in topology between the MUV datasets, differences in VS performance between the datasets are largely independent of simple molecular properties. Figure 3.8 illustrates the effect of MUV design on the topology of datasets visualized in more detail for two example datasets (AID 548, AID 858). Graphs of the nearest-neighbor function $G(t)$ (red) and the empty space function $F(t)$ (blue) provide information about the self-similarity in the set of actives and the separation between actives and decoys before (Figure 3.8A) and after (Figure 3.8B) MUV design. Here, an early ascent in the cumulative distribution of nearest neighbor distances $G(t)$, i.e. the presence of many actives with a very small distance t to their nearest neighbor, indicates high self-similarity among actives. On the other hand a late ascent in $F(t)$ implies the presence of a high proportion of decoys with a large distance t to the nearest active, i.e. a high level of separation. Any rightward shift of $F(t)$ relative to $G(t)$ results in negative values of ΣS and thereby indicates dataset clumping. The objective of MUV design is two-fold: (i) minimize the differences in dataset clumping between the datasets, i.e. generate similar curves of $G(t)$ and $F(t)$ for all datasets. (ii) Generate spatially random topologies ($\Sigma S \approx 0$), i.e. minimize the shift between $G(t)$ and $F(t)$. Figure 3.8 shows, that both goals of MUV design are achieved for the two datasets examined. Before MUV design (Figure 3.8A) both datasets exhibit the rightward shift in $F(t)$ relative to $G(t)$ that indicates dataset clumping. Moreover, the topologies of both datasets are clearly different, with a large extent of clumping in dataset AID 548 and only mild clumping in dataset AID 858. After MUV design (Figure 3.8B) both datasets show spatial randomness, i.e. no rightward shift between $F(t)$ and $G(t)$. Furthermore, the curves of both functions are similar for both datasets, i.e. differences in topology are minimized.

3.3.4 Application of MUV Datasets for LBVS Benchmarking

In order to demonstrate the utility of the MUV datasets, the PubChem bioactivity datasets before and after MUV design were encoded by different descriptors: the aforementioned simple descriptors and three additional classes of descriptors: SESP,[38] which is based on 2D atom pairs, MOE molecular properties descriptors[34] and MACCS structural keys,[123,131]

both of which have found extensive use in the literature for VS validation tasks.[77,132–134] Retrospective LBVS simulations were carried out on the original and on the MUV datasets. From each dataset one or ten actives, respectively, were chosen randomly as query molecules. The rest of the actives was pooled with the corresponding decoys and similarity searching was carried out. For searches with 10 query molecules, MAX-rule data fusion[47,48] was applied to the ranking. This was repeated 100 times. To minimize variability caused by the random selection of the query set, the average VS performance for the 100 query / validation set splits was measured by the mean area under the receiver operating characteristic curve ($mean(AUC_{ROC})$, Figure 3.9 and Table B.3) and the mean retrieval rate ($mean(RTR_{1\%})$, Table B.3). The $mean(AUC_{ROC})$ values obtained on the MUV datasets are summarized by Figure 3.9.

At a first glance, Figure 3.9 shows that none of the datasets exhibited a $mean(AUC_{ROC})$ considerably exceeding the random ranking expectation of 0.5, when encoded by simple descriptors. This indicates that the spatial randomness of MUV datasets in simple descriptor space is well reflected in the respective VS ranking. MOE descriptors and MACCS keys generally performed best. Similarity searching with 10 query compounds and data fusion worked better than single query searches. Moreover, the performance of SESP was found to be superior to simple descriptors only in very few cases. This is expected, since SESP is based on count statistics of atom pairs, which are highly correlated with the atom counts employed by simple descriptors.

However, the goal of this section of the study is not to determine the best descriptor or searching method. The objective of MUV dataset design is to prevent bias introduced by benchmark dataset composition from distorting validation experiments. Such benchmark dataset bias occurs, whenever the results of VS validation experiments largely depend on the simple property composition of the employed benchmark dataset, rather than the actual performance of the tested methods. Put differently, the validation of a method is affected by benchmark dataset bias, if its performance is highly correlated with the topology of the datasets in simple descriptor space. Consequently, benchmark dataset bias can be detected numerically using the correlation coefficient $\rho\left(\Sigma S_{simple}, mean(AUC_{ROC})\right)$ be-

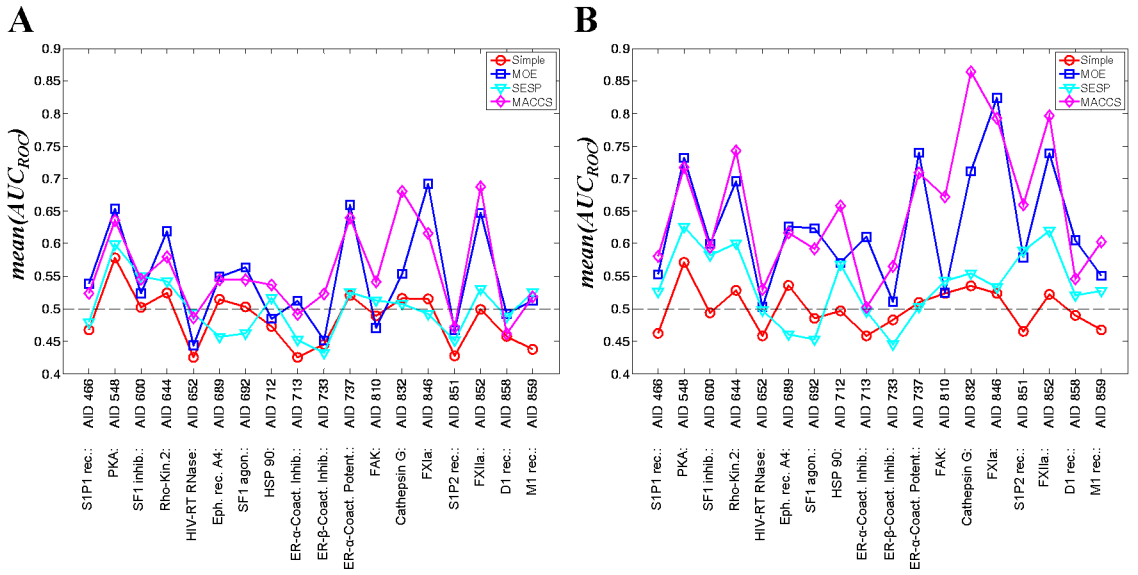


Figure 3.9: Performance of different VS methods in retrospective VS simulations on MUV datasets. (A) 1 query compound. (B) 10 query compounds. Generally, MACCS and MOE perform best. The expectation of $mean(AUC_{ROC}) = 0.5$ for random rankings is indicated by the dashed line.

tween the performance of a given VS method ($mean(AUC_{ROC})$) and dataset clumping in simple descriptor space ΣS_{simple}. Since negative values of ΣS_{simple} indicate a higher degree of clumping, VS performance and ΣS_{simple} are negatively correlated. (also see Section 2.3.2) Therefore, values of $\rho\left(\Sigma S_{simple}, mean(AUC_{ROC})\right)$ close to -1 indicate that the validation results are biased by the simple molecular properties of the benchmark datasets. Values near 0, on the other hand, indicate that the validation is not influenced by simple molecular properties. Figure 3.10 shows graphs visualizing the correlation between the $mean(AUC_{ROC})$ of the tested descriptors and ΣS_{simple} for similarity searches using 10 query compounds. A rank transformation was applied to the data, since $mean(AUC_{ROC})$ and ΣS_{simple} span considerably different numerical regions and because there is no evidence for a linear relation between them. (see Section 2.2.10) Before MUV design, a tight correlation between ΣS_{simple} and the VS performances of all tested descriptors is easily observable, both in the original data domain and after the rank transformation. Thus, before MUV design, dataset composition regarding simple molecular properties dominates the validation results of all descriptors, which means that the original datasets are affected by benchmark dataset bias. After MUV design, this correlation no longer exists

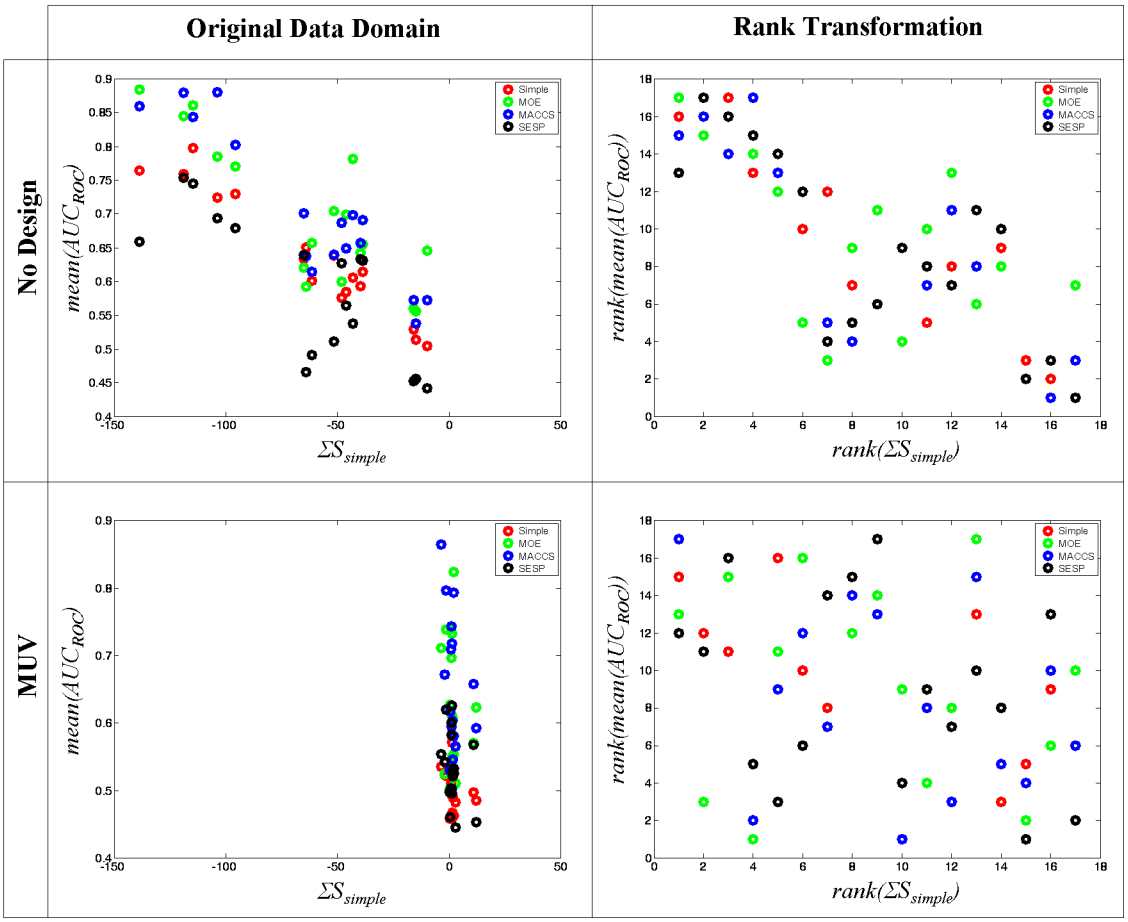


Figure 3.10: Plots of VS performance ($mean(AUC_{ROC})$) obtained with four different descriptors (Simple, MOE, MACCS, SESP) on the 18 benchmark datasets vs. dataset clumping in simple descriptor space (ΣS_{simple}) before and after MUV design (similarity searches, 10 query compounds). Before MUV design (top) a clear correlation between $mean(AUC_{ROC})$ and ΣS_{simple} is observable, both in the original data domain (left) and after a rank transformation (right). MUV design (bottom) causes a de-correlation of $mean(AUC_{ROC})$ and ΣS_{simple}. Benchmark dataset bias is prevented.

and benchmark dataset bias is effectively prevented. These observations are supported by the results shown in Table 3.5, which provides the respective spearman rank correlation coefficients $\rho\left(\Sigma S_{simple}, mean(AUC_{ROC})\right)$ for single and multiple query searches. On the original datasets all tested descriptors exhibit large negative correlation coefficients that indicate considerable benchmark dataset bias. MUV design reduces this correlation below the level of statistical significance and thereby prevents benchmark dataset bias. Here, it might be noteworthy, that in all tested cases, MOE descriptors exhibited the smallest extent of correlation with ΣS_{simple}. These results suggest that of the descriptors tested here, MOE descriptors are least susceptible to benchmark dataset bias.

Table 3.5: Correlation Coefficients Between ΣS_{simple} and $mean(AUC_{ROC})$ of Tested Descriptors.

$\rho\left(\Sigma S_{simple}, mean(AUC_{ROC})\right)^a$	No Design				MUV			
	Simple	MOE	SESP	MACCS	Simple	MOE	SESP	MACCS
1 query cmpd.	-0.84	-0.43	-0.72	-0.80	-0.22	-0.08	-0.37	-0.41
10 query cmpds.	-0.91	-0.69	-0.78	-0.79	-0.41	-0.17	-0.32	-0.41
Conf. Itv. Boundary[b]	-0.41							

[a] $n = 17$

[b] One-sided, 95%. Calculated using permutation testing. See Section 2.2.10.

3.3.5 Comparison of MUV with DUD

The directory of useful decoys DUD is a collection of VS benchmark datasets designed to prevent artificial enrichment in the validation of structure based virtual screening methods.[67] Since its publication, DUD has become the *de facto* standard for the validation of docking methods. Briefly, DUD comprises a collection of 40 datasets of actives with differing sizes compiled from various sources. All compounds in the ZINC database[71] not described to be active against one of the 40 targets were used as potential decoys. Potential decoys were required to have a Tanimoto similarity of less than 0.9 to any of the actives based on CACTVS type 2 substructure keys[135] in order to prevent the selection of compounds, that could turn out to be active if tested. Using Schrödinger QikProp[136] a vector of physical properties quite similar to the simple descriptors used here was calculated for actives and potential decoys. Based on the QikProp descriptors, the 36 most similar decoys were selected for each active from the set of potential decoys in order to generate decoy sets with minimum separation.

In order to compare the DUD collection of benchmark datasets with MUV, SD-files[80] of the DUD dataset (Release 2) were downloaded from the DUD web-site and encoded by simple and MOE descriptors. MOE descriptors were chosen, because they were found to be least susceptible to benchmark dataset bias. (see Section 3.3.4) Spatial statistics analysis and retrospective LBVS simulations were conducted in an analogous fashion as for the MUV datasets. The results are summarized in Figures 3.11 and 3.12 and Table 3.7. Each DUD dataset consists of a set of actives and a set of decoys designed for these

Table 3.6: Spatial Statistics Figures for DUD Datasets

Target	ΣS	ΣG	ΣF
ACE	-117.34	383.51	266.17
AChE	-106.75	448.56	341.81
ADA	-159.07	393.26	234.19
ALR2	-136.82	343.12	206.30
AmpC	-177.32	370.29	192.96
AR	-319.61	422.62	103.02
CDK2	8.66	327.94	336.60
COMT	-195.89	365.18	169.29
COX1	-94.74	374.16	279.42
COX2	-117.17	420.67	303.50
DHFR	-152.67	460.08	307.41
EGFR	-124.37	453.03	328.67
ER-agon.	-195.24	439.97	244.73
ER-ant.	-133.68	447.56	313.89
FGFR1	-254.40	434.20	179.80
Fxa	-137.26	395.29	258.03
GART	-302.17	457.90	155.73
GPB	-260.23	406.86	146.63
GR	-241.64	448.53	206.89
HIVPR	-135.16	368.15	232.99

Target	ΣS	ΣG	ΣF
HIVRT	-96.50	331.03	234.52
HGMR	-261.75	425.57	163.83
HSP90	-196.56	419.71	223.15
INHA	-163.98	396.59	232.61
MR	-230.41	365.93	135.53
NA	-182.47	410.04	227.57
P38	-141.92	442.53	300.61
PARP	-84.35	414.48	330.13
PDE5	-121.10	341.96	220.86
PDGFRB	-142.92	447.58	304.66
PNP	-28.43	327.00	298.57
PPARg	-154.26	412.38	258.12
PR	-126.09	412.37	286.28
RXRa	-223.82	419.50	195.68
SAHH	-214.78	455.39	240.62
SRC	-189.37	420.51	231.14
THROMBIN	-148.49	399.88	251.38
TK	-142.17	362.64	220.47
TRYPSIN	-152.76	381.95	229.19
VEGFR2	-63.31	341.35	278.04

actives. As opposed to the union of all DUD decoys, these decoy sets are termed "own" decoys by the DUD authors. These datasets of actives and the respective "own" decoys, were used here.

Most DUD datasets exhibit high levels of clumping in simple descriptor space, indicated by large negative values of ΣS_{simple} (Figure 3.11). Considerable differences in dataset topology exist between the datasets. This corresponds with significant correlation between the VS performance of both simple and MOE descriptors with ΣS_{simple} (Figure 3.12, Table 3.7). The large number of datapoints with $mean(AUC_{ROC}) \geqq 0.8$ for simple descriptors (Figure 3.12A) is particularly striking. It indicates that retrieval of actives from the DUD datasets is not very challenging, even for simple descriptors that do not encode any type of higher level molecular information like connectivity or substructure features.

Apparently, DUD is subject to considerable benchmark dataset bias. Although a detailed analysis of the underlying reasons goes beyond the scope of this study, two main

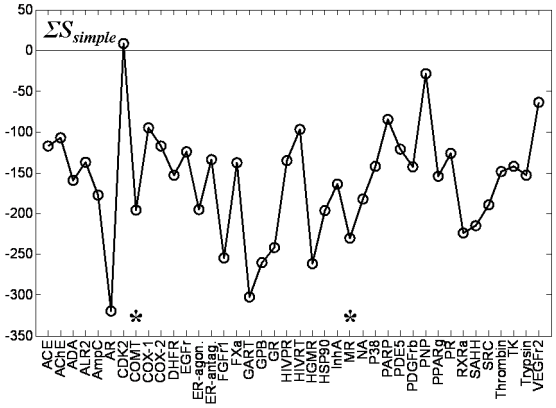


Figure 3.11: Dataset clumping of DUD benchmark datasets measured by ΣS in simple descriptor space. Large negative values indicate considerable self-similarity and separation from the decoys regarding simple molecular properties. Asterisks indicate datasets with less than 20 compounds, for which results of VS runs using 10 query compounds are not representative.

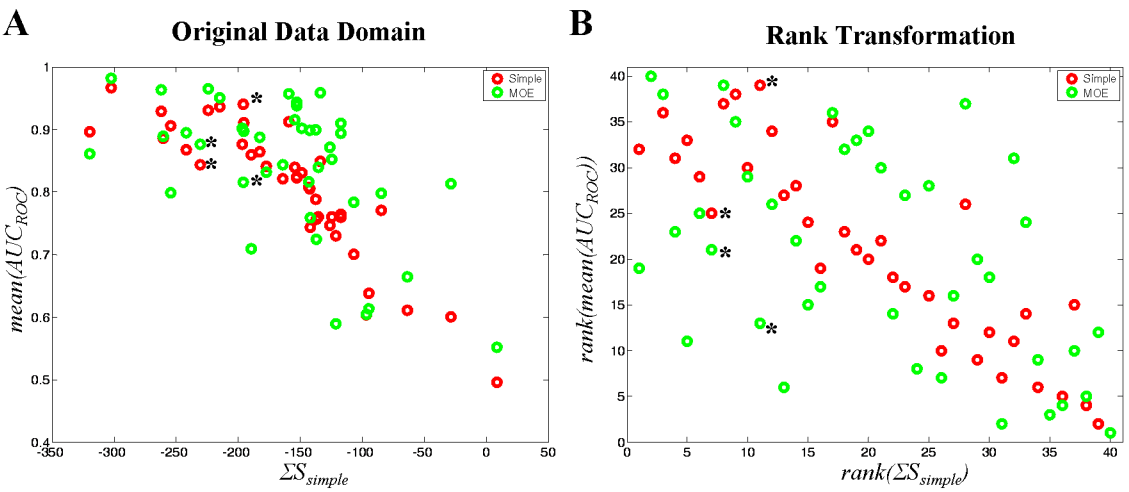


Figure 3.12: $(mean(AUC_{ROC}))$ plotted against ΣS_{simple} for MOE and simple descriptors (10 query cmpds.) on DUD benchmark datasets. (A) Original data domain. A tight correlation is observable for both simple and MOE descriptors in the original data domain. (B) The correlation persists also after rank transformation, although less obvious for MOE descriptors. Asterisks indicate datasets with less than 20 compounds.

Table 3.7: Correlation Coefficients of DUD Dataset Clumping in Simple Descriptor Space (ΣS_{simple}) and VS Performance of Tested Descriptors

$\rho\left(\Sigma S_{simple}, mean(AUC_{ROC})\right)$	Simple	MOE
10 query cmpds.	-0.92	-0.54
Significance[a]		-0.27

[a] $n = 40$

[b] One-sided, 95%. Calculated using permutation testing. See Section 2.2.10.

factors can be identified. (i) As shown recently by Good et al., DUD is seriously affected by analogue bias.[?] Applying the same algorithm by Bemis and Murcko as in the analysis of MUV datasets (see Section 3.3.2), the average number of compounds per scaffold class in DUD was determined to be 4.56 (MUV: 1.16). This is in accordance with high values of ΣG observed for all DUD datasets of actives (Table 3.6), indicating high levels of self-similarity within the datasets. Here, it is important to point out, that DUD was designed for the validation of docking programs and its application to LBVS is strongly discouraged by the DUD authors.[67,137] Consequently, no diversity selection was applied to the datasets of actives in the generation of DUD, which explains the high number of compounds per scaffold class. Since docking tools are less sensitive to the self-similarity in the dataset of actives, this might be a valid approach. However, for the validation of methods that incorporate any form of ligand information, as for instance recently by Reid et al.,[52] the analogue bias present in DUD is a critical factor. (ii) DUD decoy sets also exhibit considerable levels of separation from the actives in simple descriptor space, indicated by small values in ΣF (Table 3.6). This is surprising, because the explicit principle of DUD design is the selection of decoys that are minimally separated from the actives. Since we do not have access to the original DUD potential decoy set, it is difficult to precisely specify causes for this problem. One possible reason might be that, according to our analysis, on average 13% of the actives in DUD are inadequately embedded in decoys, with peak values of 60% for the dataset of PDGFrb inhibitors and 51% for InhA inhibitors. It is quite probable that this causes a certain degree of separation. Another possible reason is that the Tanimoto dissimilarity criterion applied to potential decoys in the design of DUD is too harsh. This might create "bubbles" devoid of decoys in chemical space around actives. MUV datasets utilize decoys, for which inactivity against the respective targets is experimentally determined, which renders a minimum dissimilarity between actives and decoys obsolete. The associated problems in the design of decoy datasets are thus circumvented.

3.4 Summary

A collection of benchmark datasets for ligand based virtual screening methods was generated from PubChem bioactivity data. These datasets minimize the influence of benchmark dataset bias on validation results and therefore provide a publicly available tool for the Maximum Unbiased Validation (MUV) of virtual screening methods. MUV is the first collection of VS benchmark datasets featuring experimentally validated decoys and incorporating the problem of chemical space embedding of actives in its design approach. The MUV datasets specifically address the validation of ligand based virtual screening techniques. Therefore both components of benchmark dataset bias, i.e. analogue bias and artificial enrichment, are minimized in the MUV datasets. With these properties, however, MUV datasets also fulfill the criteria postulated by Verdonk et al.[72] for the unbiased benchmarking of molecular docking methods. Three dimensional structures are available from the PDB for seven of the MUV targets (PKA, SF1, HIV RT-RNase, HSP90, FAK, Cathepsin G, FXIa). The respective datasets are readily applicable to the validation of SBVS methods. Furthermore, it can be safely expected that the number of MUV datasets with associated 3-D protein target structures will rise, given the rapid growth of both PubChem and the PDB. Hence, MUV is a collection of benchmark datasets that is equally unbiased for SBVS and LBVS methods. This might constitute an important progress towards comparing the performance of docking programs with ligand based VS techniques in an unbiased manner.

A workflow is presented that allows the generation of spatially optimized benchmark datasets from raw bioactivity data. As a special benefit, this workflow provides a data centered approach to detect HTS assay artifacts. The workflow is readily applicable to custom datasets of prospective users. Thereby users can generate MUV datasets customized to their specific virtual screening problems from their own in-house bioactivity data. Moreover, the filters implemented in the workflow can also be used to purge datasets for applications other than VS validation from potential unspecific binders. The workflow is modular and easily extendable regarding two important aspects: (i) New bioactivity datasets deposited in PubChem can readily be fed into the workflow, filtered for assay ar-

tifacts and inadequately embedded actives, optimized spatially and thus be integrated into the collection of MUV benchmark datasets. (ii) Because of the data centered nature of the assay artifacts filter, new experimental screening and profiling data for assay specificity can quickly be integrated. Thus, driven by the fast growth of screening data in PubChem, the MUV dataset collection will continuously be extended with new targets and datasets. Much the same way, the efficiency of the assay artifacts filter will be considerably augmented by the rapidly accumulating knowledge available in the PubChem database. The MUV collection of VS benchmark datasets (also see Appendix C) and a MATLAB[90] toolbox for the spatial statistics analysis of chemical datasets (also see Appendix D) are available on the enclosed CD-ROM or can be downloaded from the Baumann group's website: http://www.pharmchem.tu bs.de/lehre/baumann/MUV.html.

Chapter 4

Conclusion and Outlook

The research presented in this study represents the first application of Refined Nearest Neighbor Analysis in the field of chemoinformatics. Specifically, spatial statistics methods were utilized to get a deeper understanding on how the composition of benchmark datasets influences the results of LBVS validation experiments. (Chapter 2) In the course of the investigations, the topology of benchmark datasets in chemical space was identified as a key factor affecting the measured virtual screening performance. Analogue bias and artificial enrichment, which have previously been described as major factors affecting validation results, can now be quantified in terms of dataset topology. It was shown that ligand based virtual screening is affected by both phenomena. The term "*benchmark dataset bias*" was introduced for this combined effect. Based on these findings, general criteria for the prevention of benchmark dataset bias were developed. These guidelines were implemented with a workflow for the design of Maximum Unbiased Validation datasets for virtual screening. (Chapter 3) The resulting collection of unbiased benchmark datasets is based on public data and is therefore freely available to the scientific community.

Several interesting lines of research emerge from this work. This includes the adaption of the framework of spatial statistics to chemoinformatics purposes, which can now be applied to the analysis of all kinds of chemical datasets. An obvious future use certainly is the generation of unbiased datasets for suitability testing, i.e. to figure out the best VS protocol for a particular application, complementing the MUV collection of unbiased datasets for benchmarking. But there are promising applications outside the discipline of

VS validation, such as the field of library design, which seeks to optimize the chemical compound repositories of pharmaceutical companies in terms of a maximum coverage of chemical space. Here, Refined Nearest Neighbor Analysis, but also other spatial statistics methodologies might be utilized in order to quantify the complementarity of a vendor library with an in-house repository or to assess the chemical space coverage of a medicinal chemistry series of derivatives. Other, more ambitious projects might also investigate the inter-relations between the biological similarity of two pharmaceutical targets with the spatial proximity or distance of their activity spaces, thereby applying spatial statistics to the field of chemogenomics.[138]

The approach to identify frequent hitters based on PubChem data mining introduced in Section 3.2.5 also represents a promising starting point for future research. It constitutes the first completely data centered, prediction free method for the detection of frequent hitters in the public domain and can be readily applied to all kinds of chemical datasets. Moreover an *FoH* based analysis of PubChem can provide a large scale, high confidence dataset of frequent hitters, which may provide a valuable tool for the validation of current and future frequent hitter prediction algorithms.

One result that has not yet been extensively discussed in this study is the fact that current LBVS approaches are apparently unable to generate consistently high retrieval of actives in the absence of benchmark dataset bias. (see Figure 3.9, Table B.3) This might explain to some extent the fact, that very few successful applications of LBVS methods in prospective VS campaigns have yet been reported. As a consequence, the development of more powerful and accurate tools for both, ligand and structure based virtual screening poses a major challenge for the chemoinformatics community in the near future. Validation protocols providing maximum information about weaknesses or strengths of future methods will be central to this task. In the editorial of a recent special issue on the validation of virtual screening techniques of the *Journal of Computer-Aided Molecular Design*, Ajay Jain and Anthony Nicholls named three key areas on which future validation experiments must improve: data sharing, dataset preparation and reporting of results.[139] The MUV datasets presented in Chapter 3 constitute an important advance for VS validation

methods with regard to both data sharing and dataset preparation. As stated above, the MUV collection of datasets is publicly available and can easily be shared by scientists conducting validation experiments. The MUV workflow and the spatial statistics methodology presented in this study address the problem of dataset preparation with minimum dataset induced topological bias. PubChem based services, such as the PubChem Standardization Service,[140] ensure that all MUV datasets are prepared according to a common set of standards regarding tautomers, ionization, counter-ions etc. Summarizing, the research presented in the course of this study addresses two of the key issues for VS validation experiments as stated by Jain and Nicholls. Moreover, the results presented here also have implications for the future reporting of VS validation results. Given the results of this study, it is evident that the clumpiness of the employed datasets should be reported with every VS validation result. Efforts are under way in our laboratory to provide an open, community driven platform for the standardized reporting of VS validation results and the topological analysis of the employed benchmark datasets.

Appendix A

Supplementary Tables: Chapter 2

Table A.1: ΣS for all sub-samples in MOE descriptor space.

D-optimum Design						
k	50	100	150	200	250	300
ACE inhibitors	-121.3	-337.6	-421.6	-447.7	-482.1	-514.6
AChE Inhibitors	-39.2	-172.2	-244.1	-305.2	-333.8	-369.3
Angio. R. Blockers	-24.7	-151.8	-223.2	-271.6	-304.9	-334.3
COX inhibitors	-36.6	-132.3	-209.2	-250.3	-271.4	-295.5
D2 antagonists	-126.0	-230.8	-307.3	-350.9	-392.4	-419.9
HIV P. inhibitors	35.9	-76.6	-189.5	-249.8	-278.6	-316.5
5HT1A agonists	-149.1	-219.3	-254.1	-303.3	-341.5	-375.3
5HT3 antagonists	-170.4	-239.4	-307.9	-342.0	-380.4	-400.7
5HT reup. inhibitors	-189.6	-264.5	-346.3	-408.5	-441.5	-460.3
PKC inhibitors	49.7	-113.2	-225.4	-306.6	-333.2	-351.4
Renin inhibitors	-96.1	-233.9	-303.1	-359.3	-411.7	-445.2
Subst. P inhibitors	36.3	-113.9	-196.5	-243.7	-282.0	-313.3
Thrombin inhibitors	35.0	-112.8	-204.4	-256.9	-309.0	-336.6

Table A.1: ΣS for all sub-samples in MOE descriptor space.

Onion Design

k	50	100	150	200	250	300
ACE inhibitors	-337.6	-338.8	-360.0	-446.0	-494.6	-489.8
AChE Inhibitors	-230.3	-286.2	-263.5	-291.3	-304.3	-316.5
Angio. R. Blockers	-223.3	-288.8	-312.0	-309.8	-363.8	-332.7
COX inhibitors	-196.0	-218.4	-225.4	-237.5	-272.6	-276.3
D2 antagonists	-304.3	-348.4	-346.4	-346.7	-425.6	-397.1
HIV P. inhibitors	-238.6	-259.2	-275.7	-260.7	-294.2	-305.2
5HT1A agonists	-314.8	-312.3	-340.1	-356.2	-359.8	-358.8
5HT3 antagonists	-351.5	-392.9	-372.2	-392.7	-400.3	-409.6
5HT reup. inhibitors	-306.4	-347.8	-376.6	-409.9	-450.5	-453.8
PKC inhibitors	-152.0	-213.0	-227.7	-245.1	-347.9	-303.4
Renin inhibitors	-454.3	-489.3	-443.4	-426.5	-432.2	-444.7
Subst. P inhibitors	-224.2	-230.1	-234.5	-214.0	-234.2	-253.2
Thrombin inhibitors	-250.8	-274.3	-268.3	-284.9	-289.3	-301.7

Table A.1: ΣS for all sub-samples in MOE descriptor space.

Minimum Diversity Design

k	50	100	150	200	250	300
ACE inhibitors	-694.6	-665.4	-631.0	-612.6	-580.8	-560.4
AChE Inhibitors	-748.9	-673.3	-614.9	-585.8	-572.7	-557.9
Angio. R. Blockers	-925.8	-826.8	-783.0	-765.0	-743.2	-732.2
COX inhibitors	-599.8	-511.6	-484.6	-471.6	-473.8	-462.5
D2 antagonists	-691.9	-619.8	-586.9	-562.1	-543.3	-524.7
HIV P. inhibitors	-886.7	-766.5	-799.3	-747.4	-734.0	-721.1
5HT1A agonists	-706.5	-654.6	-632.8	-610.8	-607.4	-601.6
5HT3 antagonists	-747.2	-700.8	-704.6	-685.6	-660.6	-655.8
5HT reup. inhibitors	-643.9	-640.7	-575.5	-554.7	-523.9	-510.8
PKC inhibitors	-687.6	-605.9	-565.1	-548.4	-501.4	-459.5
Renin inhibitors	-968.1	-931.1	-907.7	-883.7	-863.3	-853.6
Subst. P inhibitors	-763.7	-689.4	-655.1	-620.5	-579.3	-568.0
Thrombin inhibitors	-849.3	-779.1	-696.1	-656.6	-642.5	-617.2

Table A.2: ΣS for all sub-samples in simple descriptor space.

D-optimum Design

k	50	100	150	200	250	300
ACE inhibitors	-160.9	-273.6	-290.9	-293.6	-303.2	-309.3
AChE Inhibitors	-267.8	-308.8	-328.4	-333.0	-335.1	-351.3
Angio. R. Blockers	-370.8	-397.0	-399.9	-380.9	-401.6	-416.9
COX inhibitors	-319.4	-332.0	-335.1	-340.8	-345.8	-356.8
D2 antagonists	-351.6	-370.1	-399.6	-385.1	-393.1	-394.8
HIV P. inhibitors	-222.5	-305.7	-267.5	-286.0	-307.7	-326.9
5HT1A agonists	-347.0	-357.8	-378.8	-381.7	-390.3	-397.1
5HT3 antagonists	-443.6	-492.5	-504.8	-524.1	-500.9	-502.0
5HT reup. inhibitors	-362.4	-356.3	-375.0	-381.8	-410.8	-416.0
PKC inhibitors	-117.8	-132.5	-181.1	-207.0	-225.6	-252.5
Renin inhibitors	-530.4	-583.8	-633.1	-649.8	-666.3	-679.4
Subst. P inhibitors	-156.2	-158.5	-216.3	-235.3	-260.4	-275.7
Thrombin inhibitors	-265.1	-262.9	-280.4	-290.6	-314.9	-320.0

Table A.2: Σ*S* for all sub-samples in simple descriptor space.

Onion Design

k	50	100	150	200	250	300
ACE inhibitors	-194.3	-286.5	-262.4	-288.0	-284.1	-316.9
AChE Inhibitors	-339.2	-361.3	-371.3	-348.0	-344.4	-345.6
Angio. R. Blockers	-372.5	-419.7	-398.9	-410.2	-450.2	-419.1
COX inhibitors	-412.2	-328.6	-372.3	-349.9	-360.4	-363.7
D2 antagonists	-436.2	-381.5	-404.0	-413.0	-434.6	-411.3
HIV P. inhibitors	-297.6	-272.7	-304.8	-304.8	-339.3	-307.1
5HT1A agonists	-398.1	-391.1	-384.7	-383.6	-383.6	-403.7
5HT3 antagonists	-565.8	-500.7	-488.5	-476.2	-488.8	-480.9
5HT reup. inhibitors	-364.9	-405.1	-417.6	-413.5	-404.6	-421.4
PKC inhibitors	-153.4	-238.0	-236.7	-216.4	-249.8	-251.3
Renin inhibitors	-548.2	-608.9	-671.8	-642.6	-694.5	-679.3
Subst. P inhibitors	-294.5	-264.4	-303.3	-280.3	-281.5	-289.5
Thrombin inhibitors	-344.7	-345.3	-339.5	-375.6	-310.5	-316.0

Table A.2: ΣS for all sub-samples in simple descriptor space.

Minimum Diversity Design

k	50	100	150	200	250	300
ACE inhibitors	-319.6	-348.8	-319.6	-320.5	-319.2	k
AChE Inhibitors	-451.8	-386.2	-431.4	-393.8	-396.7	-402.2
Angio. R. Blockers	-694.1	-489.5	-502.6	-493.6	-484.3	-482.9
COX inhibitors	-524.8	-423.1	-432.5	-417.6	-403.6	-403.2
D2 antagonists	-524.5	-444.8	-444.2	-455.9	-450.3	-446.8
HIV P. inhibitors	-597.2	-567.9	-502.6	-451.7	-434.6	-438.0
5HT1A agonists	-498.4	-485.0	-446.5	-446.9	-435.6	-437.2
5HT3 antagonists	-503.2	-528.0	-512.8	-511.4	-500.0	-502.4
5HT reup. inhibitors	-496.1	-502.8	-468.7	-452.5	-445.2	-450.7
PKC inhibitors	-389.6	-328.9	-353.6	-351.4	-347.5	-331.7
Renin inhibitors	-690.3	-722.4	-745.2	-762.4	-761.9	-770.4
Subst. P inhibitors	-714.6	-659.9	-564.0	-547.9	-530.9	-521.3
Thrombin inhibitors	-461.0	-513.0	-432.7	-420.1	-424.8	-364.5

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
D-Opt. Design $k = 50$												
ACE inhibitors	0.06	0.0346	0.0012	0.0016	0.0004	0.56	0.0434	0.00188542	0.00272387	0.00083846	0.07	0.58
AChE Inhibitors	0.00	0.0077	0.0001	0.0001	0.0000	0.57	0.0215	0.00046274	0.00087050	0.00040776	0.08	0.73
Angio. R. Blockers	0.01	0.0228	0.0005	0.0006	0.0001	0.50	0.0819	0.00670022	0.00717879	0.00047857	0.18	0.82
COX inhibitors	0.01	0.0138	0.0002	0.0003	0.0001	0.55	0.0313	0.00097924	0.00142867	0.00044943	0.09	0.81
D2 antagonists	0.02	0.0139	0.0002	0.0003	0.0001	0.69	0.0199	0.00039661	0.00078683	0.00039022	0.12	0.81
HIV P. inhibitors	0.01	0.0176	0.0003	0.0004	0.0000	0.43	0.1314	0.01726312	0.01767817	0.00041505	0.07	0.67
5HT1A agonists	0.02	0.0156	0.0002	0.0004	0.0001	0.68	0.0240	0.00057529	0.00092912	0.00035383	0.12	0.81
5HT3 antagonists	0.02	0.0129	0.0002	0.0003	0.0001	0.74	0.0186	0.00034567	0.00062965	0.00028398	0.25	0.91
5HT reup. inhibitors	0.05	0.0266	0.0007	0.0010	0.0003	0.72	0.0270	0.00072867	0.00102555	0.00029688	0.19	0.77
PKC inhibitors	0.01	0.0160	0.0003	0.0003	0.0001	0.34	0.0448	0.00200424	0.00250596	0.00050171	0.06	0.53
Renin inhibitors	0.06	0.0718	0.0052	0.0055	0.0003	0.59	0.1574	0.02475913	0.02526723	0.00050811	0.57	0.93
Subst. P inhibitors	0.01	0.0220	0.0005	0.0006	0.0001	0.42	0.0657	0.00431307	0.00484167	0.00052860	0.03	0.55
Thrombin inhibitors	0.02	0.0210	0.0004	0.0005	0.0001	0.47	0.1009	0.01017476	0.01067812	0.00050336	0.12	0.76

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
Onion design	$k = 50$											
ACE inhibitors	0.26	0.0432	0.0019	0.0031	0.0013	0.76	0.0329	0.00108133	0.00175339	0.00067206	0.07	0.62
AChE Inhibitors	0.10	0.0366	0.0013	0.0019	0.0006	0.75	0.0123	0.00015086	0.00061892	0.00046806	0.14	0.78
Angio. R. Blockers	0.16	0.0614	0.0038	0.0046	0.0009	0.77	0.0347	0.00120445	0.00168928	0.00048483	0.12	0.81
COX inhibitors	0.04	0.0228	0.0005	0.0008	0.0003	0.74	0.0256	0.00065290	0.00096970	0.00031680	0.09	0.85
D2 antagonists	0.11	0.0428	0.0018	0.0025	0.0007	0.83	0.0139	0.00019306	0.00043574	0.00024268	0.12	0.87
HIV P. inhibitors	0.16	0.0979	0.0096	0.0104	0.0008	0.72	0.1071	0.01146227	0.01198401	0.00052175	0.07	0.69
5HT1A agonists	0.13	0.0468	0.0022	0.0029	0.0008	0.85	0.0205	0.00041833	0.00067704	0.00025871	0.12	0.86
5HT3 antagonists	0.23	0.0615	0.0038	0.0049	0.0011	0.88	0.0000	0.00000000	0.00021873	0.00023009	0.34	0.93
5HT reup. inhibitors	0.12	0.0354	0.0013	0.0020	0.0007	0.82	0.0203	0.00041007	0.00063091	0.00022084	0.13	0.82
PKC inhibitors	0.13	0.0509	0.0026	0.0033	0.0007	0.63	0.0224	0.00050343	0.00123253	0.00072911	0.07	0.59
Renin inhibitors	0.58	0.0788	0.0062	0.0078	0.0016	0.88	0.0194	0.00037720	0.00083615	0.00045894	0.64	0.97
Subst. P inhibitors	0.12	0.0835	0.0070	0.0076	0.0007	0.73	0.0506	0.00256052	0.00307238	0.00051186	0.06	0.71
Thrombin inhibitors	0.12	0.0387	0.0015	0.0022	0.0007	0.78	0.0369	0.00135829	0.00183358	0.00047529	0.18	0.79

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
Min. Div. Des. $k = 50$												
ACE inhibitors	0.99	0.0122	0.0001	0.0002	0.0000	1.00	0.0003	0.00000007	0.00000009	0.00000002	0.13	0.75
AChE Inhibitors	1.00	0.0000	0.0000	0.0000	0.0000	1.00	0.0001	0.00000002	0.00000002	0.00000000	0.31	0.86
Angio. R. Blockers	1.00	0.0000	0.0000	0.0000	0.0000	1.00	0.0001	0.00000000	0.00000000	0.00000000	0.61	0.95
COX inhibitors	0.82	0.0810	0.0066	0.0075	0.0010	0.99	0.0031	0.00000931	0.00001008	0.00000077	0.29	0.90
D2 antagonists	0.96	0.0264	0.0007	0.0009	0.0002	1.00	0.0005	0.00000025	0.00000030	0.00000006	0.25	0.89
HIV P. inhibitors	1.00	0.0000	0.0000	0.0000	0.0000	1.00	0.0000	0.00000000	0.00000000	0.00000000	0.48	0.92
5HT1A agonists	1.00	0.0075	0.0001	0.0001	0.0000	1.00	0.0002	0.00000004	0.00000004	0.00000001	0.31	0.88
5HT3 antagonists	1.00	0.0000	0.0000	0.0000	0.0000	1.00	0.0001	0.00000001	0.00000001	0.00000000	0.28	0.92
5HT reup. inhibitors	0.86	0.0562	0.0032	0.0040	0.0008	1.00	0.0011	0.00000131	0.00000152	0.00000021	0.31	0.89
PKC inhibitors	0.99	0.0357	0.0013	0.0014	0.0001	1.00	0.0008	0.00000070	0.00000073	0.00000004	0.29	0.73
Renin inhibitors	1.00	0.0000	0.0000	0.0000	0.0000	1.00	0.0000	0.00000000	0.00000000	0.00000000	0.81	0.99
Subst. P inhibitors	1.00	0.0121	0.0001	0.0002	0.0000	1.00	0.0003	0.00000011	0.00000012	0.00000001	0.66	0.98
Thrombin inhibitors	1.00	0.0000	0.0000	0.0000	0.0000	1.00	0.0000	0.00000000	0.00000000	0.00000000	0.30	0.88

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
D-Opt. Des.	$k = 100$											
ACE inhibitors	0.12	0.0289	0.0008	0.0011	0.0003	0.60	0.0474	0.00225017	0.00260941	0.00035924	0.09	0.63
AChE Inhibitors	0.03	0.0226	0.0005	0.0006	0.0001	0.61	0.0319	0.00101844	0.00120421	0.00018578	0.07	0.75
Angio. R. Blockers	0.04	0.0240	0.0006	0.0007	0.0001	0.58	0.0779	0.00607216	0.00627618	0.00020402	0.18	0.83
COX inhibitors	0.03	0.0158	0.0003	0.0003	0.0001	0.59	0.0269	0.00072097	0.00091009	0.00018912	0.07	0.79
D2 antagonists	0.05	0.0219	0.0005	0.0006	0.0001	0.72	0.0272	0.00073803	0.00088580	0.00014777	0.12	0.80
HIV P. inhibitors	0.02	0.0192	0.0004	0.0004	0.0000	0.46	0.1100	0.01209111	0.01228463	0.00019352	0.09	0.74
5HT1A agonists	0.04	0.0168	0.0003	0.0004	0.0001	0.71	0.0239	0.00057309	0.00072725	0.00015416	0.10	0.81
5HT3 antagonists	0.05	0.0215	0.0005	0.0006	0.0001	0.77	0.0244	0.00059470	0.00071081	0.00011611	0.27	0.92
5HT reup. inhibitors	0.06	0.0274	0.0007	0.0009	0.0002	0.75	0.0281	0.00078936	0.00090458	0.00011522	0.12	0.78
PKC inhibitors	0.03	0.0176	0.0003	0.0004	0.0001	0.43	0.0421	0.00177485	0.00202707	0.00025222	0.05	0.52
Renin inhibitors	0.09	0.1075	0.0116	0.0118	0.0002	0.67	0.1420	0.02016941	0.02038481	0.00021540	0.54	0.95
Subst. P inhibitors	0.03	0.0264	0.0007	0.0008	0.0001	0.45	0.0759	0.00575920	0.00599660	0.00023740	0.02	0.55
Thrombin inhibitors	0.05	0.0408	0.0017	0.0018	0.0001	0.58	0.1008	0.01015641	0.01040041	0.00024400	0.12	0.73

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
Onion Des.		$k = 100$										
ACE inhibitors	0.24	0.0478	0.0023	0.0028	0.0005	0.76	0.0330	0.00109001	0.00137826	0.00028825	0.09	0.65
AChE Inhibitors	0.08	0.0262	0.0007	0.0009	0.0002	0.77	0.0157	0.00024545	0.00039691	0.00015147	0.11	0.77
Angio. R. Blockers	0.19	0.0619	0.0038	0.0043	0.0004	0.81	0.0362	0.00131244	0.00148617	0.00017373	0.19	0.83
COX inhibitors	0.06	0.0221	0.0005	0.0006	0.0002	0.75	0.0282	0.00079422	0.00094824	0.00015402	0.06	0.81
D2 antagonists	0.15	0.0372	0.0014	0.0017	0.0004	0.85	0.0120	0.00014418	0.00024673	0.00010255	0.12	0.82
HIV P. inhibitors	0.14	0.1024	0.0105	0.0108	0.0003	0.76	0.0754	0.00569229	0.00587799	0.00018570	0.09	0.71
5HT1A agonists	0.10	0.0361	0.0013	0.0016	0.0003	0.83	0.0165	0.00027097	0.00038006	0.00010909	0.12	0.81
5HT3 antagonists	0.26	0.0488	0.0024	0.0029	0.0006	0.90	0.0116	0.00013374	0.00019772	0.00006398	0.32	0.92
5HT reup. inhibitors	0.14	0.0452	0.0020	0.0024	0.0003	0.83	0.0207	0.00042812	0.00052608	0.00009796	0.13	0.81
PKC inhibitors	0.12	0.0396	0.0016	0.0019	0.0003	0.64	0.0189	0.00035549	0.00065529	0.00029980	0.09	0.62
Renin inhibitors	0.51	0.1034	0.0107	0.0114	0.0007	0.92	0.0222	0.00049227	0.00058565	0.00009338	0.58	0.95
Subst. P inhibitors	0.07	0.0356	0.0013	0.0015	0.0002	0.68	0.0610	0.00371864	0.00392793	0.00020928	0.05	0.71
Thrombin inhibitors	0.16	0.0715	0.0051	0.0055	0.0004	0.78	0.0471	0.00221576	0.00242902	0.00021327	0.15	0.77

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
Min. Div. Des. $k = 100$												
ACE inhibitors	0.88	0.0542	0.0029	0.0032	0.0003	1.00	0.0015	0.00000221	0.00002364	0.00000014	0.14	0.73
AChE Inhibitors	0.85	0.0574	0.0033	0.0037	0.0004	1.00	0.0014	0.00000196	0.00002136	0.00000012	0.13	0.79
Angio. R. Blockers	1.00	0.0000	0.0000	0.0000	0.0000	1.00	0.0001	0.00000002	0.00000006	0.00000001	0.22	0.87
COX inhibitors	0.38	0.0716	0.0051	0.0058	0.0007	0.96	0.0107	0.00011464	0.00010258	0.00000571	0.09	0.85
D2 antagonists	0.63	0.0610	0.0037	0.0044	0.0007	0.99	0.0022	0.00000485	0.00001628	0.00000042	0.16	0.86
HIV P. inhibitors	0.91	0.0618	0.0038	0.0040	0.0002	1.00	0.0013	0.00000157	0.00000077	0.00000007	0.31	0.82
5HT1A agonists	0.89	0.0477	0.0023	0.0026	0.0003	1.00	0.0009	0.00000079	0.00000260	0.00000007	0.26	0.87
5HT3 antagonists	0.99	0.0159	0.0003	0.0003	0.0000	1.00	0.0003	0.00000010	0.00000042	0.00000001	0.34	0.93
5HT reup. inhibitors	0.74	0.0689	0.0048	0.0053	0.0005	0.99	0.0023	0.00000530	0.00004965	0.00000035	0.26	0.89
PKC inhibitors	0.79	0.0694	0.0048	0.0053	0.0005	0.99	0.0042	0.00001782	0.00011250	0.00000094	0.19	0.71
Renin inhibitors	1.00	0.0000	0.0000	0.0000	0.0000	1.00	0.0000	0.00000000	0.00000000	0.00000000	0.83	0.99
Subst. P inhibitors	0.85	0.0653	0.0043	0.0046	0.0004	1.00	0.0017	0.00000291	0.00001412	0.00000015	0.35	0.95
Thrombin inhibitors	1.00	0.0007	0.0000	0.0000	0.0000	1.00	0.0002	0.00000003	0.00000070	0.00000000	0.30	0.88

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
D-Opt. Des. $k = 150$												
ACE inhibitors	0.15	0.0289	0.0014	0.0016	0.0002	0.66	0.0574	0.00329734	0.00350967	0.00021233	0.08	0.64
AChE Inhibitors	0.05	0.0226	0.0006	0.0007	0.0001	0.64	0.0282	0.00079684	0.00091966	0.00012283	0.10	0.75
Angio. R. Blockers	0.05	0.0240	0.0010	0.0011	0.0001	0.61	0.0787	0.00620083	0.00633161	0.00013078	0.18	0.83
COX inhibitors	0.04	0.0158	0.0003	0.0004	0.0001	0.61	0.0302	0.00091329	0.00103210	0.00011881	0.06	0.79
D2 antagonists	0.06	0.0219	0.0005	0.0006	0.0001	0.77	0.0204	0.00041744	0.00050433	0.00008689	0.11	0.82
HIV P. inhibitors	0.04	0.0192	0.0008	0.0009	0.0001	0.51	0.1040	0.01080688	0.01093469	0.00012781	0.09	0.71
5HT1A agonists	0.05	0.0168	0.0004	0.0005	0.0001	0.73	0.0221	0.00048999	0.00058000	0.00009001	0.11	0.81
5HT3 antagonists	0.08	0.0215	0.0005	0.0007	0.0001	0.79	0.0196	0.00038468	0.00045430	0.00006961	0.26	0.93
5HT reup. inhibitors	0.08	0.0274	0.0003	0.0005	0.0001	0.77	0.0254	0.00064477	0.00071842	0.00007365	0.13	0.79
PKC inhibitors	0.06	0.0176	0.0005	0.0006	0.0001	0.49	0.0458	0.00209649	0.00227637	0.00017988	0.05	0.52
Renin inhibitors	0.13	0.1075	0.0120	0.0122	0.0002	0.72	0.1296	0.01680517	0.01692903	0.00012386	0.46	0.92
Subst. P inhibitors	0.05	0.0264	0.0012	0.0013	0.0001	0.49	0.0645	0.00415871	0.00431501	0.00015630	0.03	0.58
Thrombin inhibitors	0.06	0.0408	0.0011	0.0012	0.0001	0.59	0.0881	0.00776055	0.00791522	0.00015467	0.11	0.74

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
Onion Des.	$k = 150$											
ACE inhibitors	0.24	0.0478	0.0025	0.0028	0.0003	0.76	0.0359	0.00129092	0.00147656	0.00018564	0.09	0.65
AChE Inhibitors	0.06	0.0262	0.0004	0.0005	0.0001	0.76	0.0141	0.00019793	0.00029980	0.00010187	0.11	0.79
Angio. R. Blockers	0.17	0.0619	0.0038	0.0041	0.0003	0.82	0.0473	0.00224015	0.00232599	0.00008584	0.13	0.83
COX inhibitors	0.06	0.0221	0.0005	0.0006	0.0001	0.73	0.0231	0.00053503	0.00062594	0.00009091	0.08	0.82
D2 antagonists	0.12	0.0372	0.0011	0.0013	0.0002	0.85	0.0125	0.00015513	0.00021950	0.00006437	0.13	0.83
HIV P. inhibitors	0.14	0.1024	0.0101	0.0103	0.0002	0.76	0.0952	0.00907102	0.00917928	0.00010826	0.12	0.75
5HT1A agonists	0.10	0.0361	0.0008	0.0010	0.0002	0.84	0.0139	0.00019290	0.00025638	0.00006348	0.10	0.81
5HT3 antagonists	0.23	0.0488	0.0025	0.0028	0.0003	0.90	0.0082	0.00006719	0.00010897	0.00004178	0.27	0.92
5HT reup. inhibitors	0.14	0.0452	0.0009	0.0011	0.0002	0.84	0.0156	0.00024420	0.00029732	0.00005312	0.15	0.83
PKC inhibitors	0.10	0.0396	0.0011	0.0013	0.0002	0.63	0.0288	0.00083171	0.00101262	0.00018091	0.07	0.59
Renin inhibitors	0.46	0.1034	0.0351	0.0355	0.0004	0.90	0.0449	0.00201362	0.00208902	0.00007540	0.65	0.96
Subst. P inhibitors	0.10	0.0356	0.0036	0.0038	0.0002	0.71	0.0653	0.00425937	0.00438983	0.00013045	0.06	0.70
Thrombin inhibitors	0.15	0.0715	0.0041	0.0043	0.0002	0.80	0.0465	0.00215936	0.00228704	0.00012768	0.12	0.76

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
Min. Div. Des. $k = 150$												
ACE inhibitors	0.71	0.0542	0.0037	0.0041	0.0004	0.99	0.0048	0.00002284	0.00002364	0.00000080	0.12	0.68
AChE Inhibitors	0.67	0.0574	0.0085	0.0089	0.0004	0.99	0.0046	0.00002073	0.00002136	0.00000063	0.13	0.82
Angio. R. Blockers	1.00	0.0000	0.0000	0.0000	0.0000	1.00	0.0002	0.00000005	0.00000006	0.00000001	0.24	0.88
COX inhibitors	0.30	0.0716	0.0026	0.0030	0.0004	0.95	0.0098	0.00009622	0.00010258	0.00000636	0.09	0.84
D2 antagonists	0.46	0.0610	0.0045	0.0049	0.0004	0.98	0.0039	0.00001543	0.00001628	0.00000085	0.14	0.85
HIV P. inhibitors	0.93	0.0618	0.0022	0.0023	0.0001	1.00	0.0009	0.00000074	0.00000077	0.00000003	0.26	0.84
5HT1A agonists	0.74	0.0477	0.0031	0.0035	0.0003	0.99	0.0016	0.00000242	0.00000260	0.00000017	0.19	0.85
5HT3 antagonists	0.94	0.0159	0.0012	0.0013	0.0001	1.00	0.0006	0.00000040	0.00000042	0.00000002	0.33	0.93
5HT reup. inhibitors	0.48	0.0689	0.0042	0.0046	0.0004	0.97	0.0068	0.00004600	0.00004965	0.00000365	0.24	0.87
PKC inhibitors	0.58	0.0694	0.0085	0.0090	0.0004	0.97	0.0105	0.00010922	0.00011250	0.00000329	0.15	0.69
Renin inhibitors	1.00	0.0000	0.0000	0.0000	0.0000	1.00	0.0000	0.00000000	0.00000000	0.00000000	0.79	0.99
Subst. P inhibitors	0.72	0.0653	0.0040	0.0043	0.0004	0.99	0.0037	0.00001365	0.00001412	0.00000047	0.27	0.91
Thrombin inhibitors	0.95	0.0007	0.0017	0.0018	0.0001	1.00	0.0008	0.00000068	0.00000070	0.00000002	0.25	0.85

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
D-Opt. Des. $k = 200$												
ACE inhibitors	0.18	0.0371	0.0014	0.0016	0.0002	0.70	0.0551	0.00303955	0.00319145	0.00015190	0.08	0.63
AChE Inhibitors	0.06	0.0239	0.0004	0.0005	0.0001	0.66	0.0243	0.00059266	0.00067860	0.00008594	0.09	0.75
Angio. R. Blockers	0.06	0.0314	0.0010	0.0011	0.0001	0.65	0.0747	0.00557715	0.00567237	0.00009522	0.15	0.82
COX inhibitors	0.05	0.0173	0.0003	0.0004	0.0001	0.64	0.0264	0.00069790	0.00078024	0.00008234	0.07	0.79
D2 antagonists	0.09	0.0230	0.0009	0.0010	0.0001	0.80	0.0200	0.00039954	0.00045367	0.00005413	0.10	0.82
HIV P. inhibitors	0.05	0.0286	0.0009	0.0010	0.0001	0.56	0.1152	0.01327318	0.01336583	0.00009265	0.08	0.71
5HT1A agonists	0.06	0.0199	0.0003	0.0004	0.0001	0.75	0.0185	0.00034366	0.00040276	0.00005910	0.11	0.82
5HT3 antagonists	0.09	0.0231	0.0004	0.0005	0.0001	0.81	0.0186	0.00034731	0.00039316	0.00004585	0.28	0.93
5HT reup. inhibitors	0.11	0.0182	0.0003	0.0004	0.0001	0.80	0.0273	0.00074460	0.00079304	0.00004844	0.12	0.80
PKC inhibitors	0.09	0.0224	0.0007	0.0008	0.0001	0.53	0.0404	0.00162870	0.00176363	0.00013493	0.05	0.54
Renin inhibitors	0.11	0.1094	0.0071	0.0072	0.0001	0.72	0.1117	0.01247628	0.01256707	0.00009079	0.50	0.93
Subst. P inhibitors	0.04	0.0349	0.0005	0.0006	0.0001	0.51	0.0654	0.00427405	0.00438491	0.00011085	0.03	0.58
Thrombin inhibitors	0.08	0.0334	0.0015	0.0016	0.0001	0.63	0.0840	0.00706386	0.00717715	0.00011329	0.10	0.72

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
Onion Des.	$k = 200$											
ACE inhibitors	0.32	0.0501	0.0024	0.0027	0.0003	0.82	0.0451	0.00203558	0.00214184	0.00010626	0.08	0.65
AChE Inhibitors	0.08	0.0199	0.0004	0.0005	0.0001	0.77	0.0143	0.00020519	0.00027167	0.00006648	0.08	0.76
Angio. R. Blockers	0.16	0.0620	0.0046	0.0048	0.0002	0.81	0.0463	0.00214242	0.00221533	0.00007291	0.14	0.85
COX inhibitors	0.05	0.0213	0.0002	0.0003	0.0001	0.73	0.0250	0.00062656	0.00068957	0.00006300	0.10	0.80
D2 antagonists	0.15	0.0339	0.0015	0.0016	0.0002	0.86	0.0147	0.00021605	0.00025628	0.00004023	0.13	0.84
HIV P. inhibitors	0.12	0.1003	0.0078	0.0079	0.0001	0.75	0.0963	0.00927280	0.00935274	0.00007994	0.11	0.73
5HT1A agonists	0.11	0.0289	0.0005	0.0006	0.0001	0.84	0.0132	0.00017445	0.00021954	0.00004509	0.12	0.82
5HT3 antagonists	0.23	0.0497	0.0022	0.0024	0.0002	0.90	0.0095	0.00009090	0.00012157	0.00003067	0.30	0.91
5HT reup. inhibitors	0.15	0.0301	0.0011	0.0013	0.0002	0.85	0.0263	0.00069179	0.00073252	0.00004073	0.14	0.82
PKC inhibitors	0.11	0.0332	0.0013	0.0015	0.0001	0.63	0.0272	0.00073757	0.00088492	0.00014735	0.07	0.57
Renin inhibitors	0.48	0.1875	0.0282	0.0285	0.0003	0.91	0.0457	0.00208439	0.00213210	0.00004772	0.58	0.96
Subst. P inhibitors	0.06	0.0602	0.0015	0.0016	0.0001	0.69	0.0629	0.00395472	0.00404993	0.00009522	0.05	0.69
Thrombin inhibitors	0.16	0.0641	0.0065	0.0067	0.0002	0.79	0.0500	0.00250375	0.00259342	0.00008967	0.16	0.81

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
Min. Div. Des. $k = 200$												
ACE inhibitors	0.58	0.0608	0.0037	0.0040	0.0003	0.98	0.0080	0.00006333	0.00006619	0.00000286	0.11	0.67
AChE Inhibitors	0.49	0.0923	0.0045	0.0048	0.0003	0.98	0.0055	0.00003008	0.00003136	0.00000128	0.13	0.79
Angio. R. Blockers	0.98	0.0016	0.0004	0.0004	0.0000	1.00	0.0003	0.00000010	0.00000011	0.00000001	0.23	0.88
COX inhibitors	0.22	0.0508	0.0024	0.0026	0.0002	0.93	0.0136	0.00018488	0.00019163	0.00000675	0.10	0.83
D2 antagonists	0.35	0.0669	0.0040	0.0043	0.0003	0.97	0.0070	0.00004961	0.00005109	0.00000148	0.13	0.86
HIV P. inhibitors	0.79	0.0465	0.0070	0.0072	0.0002	0.99	0.0030	0.00000923	0.00000938	0.00000014	0.19	0.80
5HT1A agonists	0.56	0.0561	0.0029	0.0032	0.0003	0.99	0.0026	0.00000657	0.00000693	0.00000036	0.16	0.83
5HT3 antagonists	0.87	0.0343	0.0019	0.0021	0.0002	1.00	0.0009	0.00000085	0.00000089	0.00000004	0.33	0.92
5HT reup. inhibitors	0.38	0.0645	0.0038	0.0041	0.0003	0.94	0.0093	0.00008578	0.00009331	0.00000754	0.20	0.85
PKC inhibitors	0.44	0.0924	0.0053	0.0056	0.0003	0.95	0.0116	0.00013405	0.00014213	0.00000808	0.14	0.69
Renin inhibitors	1.00	0.0000	0.0000	0.0000	0.0000	1.00	0.0001	0.00000001	0.00000001	0.00000000	0.76	0.98
Subst. P inhibitors	0.59	0.0630	0.0057	0.0060	0.0003	0.98	0.0062	0.00003898	0.00003984	0.00000086	0.21	0.89
Thrombin inhibitors	0.87	0.0413	0.0051	0.0052	0.0001	1.00	0.0017	0.00000278	0.00000284	0.00000006	0.21	0.83

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
D-Opt. Design $k = 250$												
ACE inhibitors	0.23	0.0370	0.0021	0.0023	0.0002	0.74	0.0539	0.00290673	0.00301699	0.00011026	0.09	0.63
AChE Inhibitors	0.07	0.0205	0.0006	0.0007	0.0001	0.68	0.0272	0.00074102	0.00080638	0.00006536	0.08	0.74
Angio. R. Blockers	0.08	0.0318	0.0020	0.0020	0.0001	0.69	0.0758	0.00575232	0.00582277	0.00007044	0.16	0.83
COX inhibitors	0.05	0.0183	0.0002	0.0003	0.0000	0.66	0.0278	0.00077361	0.00083350	0.00005989	0.06	0.80
D2 antagonists	0.12	0.0306	0.0013	0.0014	0.0001	0.82	0.0170	0.00028914	0.00032829	0.00003915	0.11	0.82
HIV P. inhibitors	0.05	0.0307	0.0013	0.0013	0.0001	0.56	0.1108	0.01227170	0.01234420	0.00007250	0.10	0.73
5HT1A agonists	0.06	0.0170	0.0003	0.0004	0.0001	0.76	0.0217	0.00047176	0.00051705	0.00004529	0.11	0.82
5HT3 antagonists	0.11	0.0208	0.0007	0.0008	0.0001	0.83	0.0188	0.00035201	0.00038600	0.00003399	0.28	0.92
5HT reup. inhibitors	0.13	0.0171	0.0008	0.0010	0.0001	0.81	0.0256	0.00065320	0.00069109	0.00003789	0.13	0.81
PKC inhibitors	0.10	0.0259	0.0007	0.0008	0.0001	0.56	0.0389	0.00151234	0.00161888	0.00010654	0.05	0.55
Renin inhibitors	0.14	0.0840	0.0124	0.0125	0.0001	0.74	0.1116	0.01244580	0.01251775	0.00007195	0.49	0.93
Subst. P inhibitors	0.06	0.0234	0.0010	0.0010	0.0001	0.54	0.0743	0.00551629	0.00560093	0.00008464	0.03	0.60
Thrombin inhibitors	0.08	0.0388	0.0010	0.0011	0.0001	0.65	0.0712	0.00506424	0.00515145	0.00008721	0.10	0.73

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
Onion Des.	$k = 250$											
ACE inhibitors	0.35	0.0489	0.0039	0.0042	0.0003	0.80	0.0431	0.00185434	0.00199421	0.00013987	0.09	0.65
AChE Inhibitors	0.09	0.0209	0.0004	0.0005	0.0001	0.79	0.0145	0.00020906	0.00025921	0.00005015	0.08	0.76
Angio. R. Blockers	0.15	0.0680	0.0028	0.0030	0.0001	0.82	0.0430	0.00185025	0.00190018	0.00004993	0.17	0.86
COX inhibitors	0.07	0.0154	0.0007	0.0008	0.0001	0.74	0.0228	0.00051845	0.00057598	0.00005753	0.09	0.81
D2 antagonists	0.18	0.0384	0.0016	0.0018	0.0002	0.88	0.0147	0.00021686	0.00024747	0.00003061	0.14	0.86
HIV P. inhibitors	0.14	0.0884	0.0076	0.0077	0.0001	0.77	0.0848	0.00719824	0.00725846	0.00006022	0.09	0.70
5HT1A agonists	0.13	0.0221	0.0007	0.0008	0.0001	0.84	0.0121	0.00014760	0.00018438	0.00003678	0.11	0.81
5HT3 antagonists	0.24	0.0469	0.0018	0.0020	0.0002	0.90	0.0068	0.00004672	0.00007168	0.00002497	0.28	0.91
5HT reup. inhibitors	0.22	0.0339	0.0024	0.0027	0.0002	0.86	0.0219	0.00047975	0.00052646	0.00004671	0.14	0.82
PKC inhibitors	0.16	0.0366	0.0020	0.0022	0.0002	0.70	0.0293	0.00086115	0.00096948	0.00010833	0.07	0.58
Renin inhibitors	0.48	0.1679	0.0310	0.0312	0.0002	0.90	0.0561	0.00314301	0.00318415	0.00004114	0.66	0.97
Subst. P inhibitors	0.08	0.0390	0.0027	0.0028	0.0001	0.69	0.0765	0.00585373	0.00592444	0.00007071	0.05	0.70
Thrombin inhibitors	0.17	0.0806	0.0067	0.0069	0.0001	0.81	0.0572	0.00327369	0.00333496	0.00006127	0.13	0.77

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
Min. Div. Des. $k = 250$												
ACE inhibitors	0.48	0.0610	0.0031	0.0034	0.0002	0.94	0.0216	0.00046691	0.00047912	0.00001221	0.09	0.66
AChE Inhibitors	0.47	0.0671	0.0047	0.0049	0.0003	0.97	0.0086	0.00007317	0.00007465	0.00000148	0.13	0.80
Angio. R. Blockers	0.93	0.0200	0.0011	0.0012	0.0001	1.00	0.0003	0.00000007	0.00000009	0.00000001	0.21	0.87
COX inhibitors	0.22	0.0490	0.0020	0.0022	0.0002	0.92	0.0128	0.00016511	0.00017210	0.00000699	0.09	0.82
D2 antagonists	0.30	0.0629	0.0032	0.0034	0.0002	0.96	0.0087	0.00007598	0.00007830	0.00000232	0.12	0.85
HIV P. inhibitors	0.71	0.0835	0.0086	0.0088	0.0002	0.99	0.0045	0.00002062	0.00002083	0.00000021	0.17	0.78
5HT1A agonists	0.50	0.0536	0.0029	0.0032	0.0003	0.98	0.0032	0.00001022	0.00001058	0.00000036	0.14	0.83
5HT3 antagonists	0.77	0.0441	0.0037	0.0039	0.0002	0.99	0.0018	0.00000326	0.00000335	0.00000009	0.30	0.92
5HT reup. inhibitors	0.28	0.0620	0.0026	0.0028	0.0002	0.92	0.0108	0.00011765	0.00012747	0.00000982	0.17	0.84
PKC inhibitors	0.34	0.0728	0.0060	0.0063	0.0002	0.90	0.0198	0.00039108	0.00040815	0.00001707	0.11	0.68
Renin inhibitors	1.00	0.0000	0.0000	0.0000	0.0000	1.00	0.0001	0.00000002	0.00000002	0.00000000	0.76	0.98
Subst. P inhibitors	0.48	0.0755	0.0041	0.0044	0.0002	0.97	0.0085	0.00007223	0.00007454	0.00000231	0.19	0.89
Thrombin inhibitors	0.77	0.0711	0.0063	0.0065	0.0002	0.99	0.0030	0.00000894	0.00000908	0.00000014	0.22	0.83

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
D-Opt. Des. $k = 300$												
ACE inhibitors	0.26	0.0456	0.0019	0.0021	0.0002	0.76	0.0458	0.00209809	0.00218962	0.00009153	0.09	0.64
AChE Inhibitors	0.08	0.0251	0.0006	0.0007	0.0001	0.70	0.0239	0.00057041	0.00062388	0.00005346	0.08	0.74
Angio. R. Blockers	0.09	0.0442	0.0019	0.0019	0.0001	0.71	0.0767	0.00587782	0.00593557	0.00005775	0.17	0.84
COX inhibitors	0.06	0.0155	0.0004	0.0004	0.0000	0.68	0.0237	0.00056005	0.00060894	0.00004889	0.08	0.81
D2 antagonists	0.14	0.0361	0.0018	0.0019	0.0001	0.84	0.0133	0.00017725	0.00020681	0.00002957	0.11	0.82
HIV P. inhibitors	0.06	0.0359	0.0009	0.0010	0.0001	0.60	0.1040	0.01082398	0.01088338	0.00005940	0.09	0.72
5HT1A agonists	0.07	0.0181	0.0005	0.0005	0.0001	0.78	0.0205	0.00042105	0.00045482	0.00003377	0.12	0.82
5HT3 antagonists	0.12	0.0270	0.0007	0.0008	0.0001	0.84	0.0184	0.00033787	0.00036269	0.00002481	0.28	0.92
5HT reup. inhibitors	0.16	0.0288	0.0008	0.0009	0.0001	0.84	0.0195	0.00037833	0.00040546	0.00002714	0.14	0.82
PKC inhibitors	0.10	0.0259	0.0008	0.0009	0.0001	0.59	0.0411	0.00169064	0.00177776	0.00008711	0.06	0.55
Renin inhibitors	0.19	0.1115	0.0171	0.0172	0.0001	0.79	0.1085	0.01177901	0.01182956	0.00005056	0.49	0.94
Subst. P inhibitors	0.07	0.0312	0.0016	0.0016	0.0001	0.57	0.0706	0.00497884	0.00505026	0.00007143	0.04	0.62
Thrombin inhibitors	0.09	0.0321	0.0013	0.0014	0.0001	0.67	0.0672	0.00451275	0.00458393	0.00007118	0.12	0.75

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
Onion Des.	$k = 300$											
ACE inhibitors	0.32	0.0624	0.0021	0.0023	0.0002	0.78	0.0398	0.00158501	0.00167292	0.00008790	0.09	0.64
AChE Inhibitors	0.09	0.0204	0.0006	0.0006	0.0001	0.79	0.0151	0.00022718	0.00026992	0.00004274	0.08	0.76
Angio. R. Blockers	0.17	0.0534	0.0038	0.0039	0.0001	0.82	0.0360	0.00129505	0.00133864	0.00004358	0.15	0.84
COX inhibitors	0.06	0.0263	0.0003	0.0004	0.0001	0.75	0.0252	0.00063733	0.00067830	0.00004097	0.08	0.81
D2 antagonists	0.16	0.0394	0.0014	0.0015	0.0001	0.86	0.0128	0.00016374	0.00019079	0.00002705	0.11	0.83
HIV P. inhibitors	0.13	0.0873	0.0070	0.0071	0.0001	0.75	0.0959	0.00918977	0.00924022	0.00005046	0.10	0.73
5HT1A agonists	0.12	0.0266	0.0007	0.0008	0.0001	0.85	0.0160	0.00025488	0.00028198	0.00002710	0.12	0.82
5HT3 antagonists	0.24	0.0426	0.0024	0.0025	0.0002	0.90	0.0107	0.00011439	0.00013284	0.00001845	0.29	0.92
5HT reup. inhibitors	0.18	0.0495	0.0012	0.0013	0.0001	0.85	0.0219	0.00048061	0.00050609	0.00002547	0.14	0.82
PKC inhibitors	0.15	0.0450	0.0021	0.0022	0.0001	0.66	0.0281	0.00079130	0.00087905	0.00008775	0.07	0.57
Renin inhibitors	0.44	0.1761	0.0390	0.0392	0.0002	0.90	0.0569	0.00324273	0.00327734	0.00003461	0.55	0.95
Subst. P inhibitors	0.08	0.0523	0.0022	0.0023	0.0001	0.70	0.0702	0.00492756	0.00498563	0.00005807	0.05	0.71
Thrombin inhibitors	0.16	0.0820	0.0057	0.0058	0.0001	0.79	0.0494	0.00244499	0.00249905	0.00005406	0.11	0.76

Table A.3: VS Figures of Merit for dataset sub-samples.

	MOE										Simple	
	$mean(RTR_{1\%})$	$\sigma_{top}(RTR_{1\%})$	$\sigma^2_{top}(RTR_{1\%})$	$\sigma^2(RTR_{1\%})$	$\sigma^2_B(RTR_{1\%})$	$mean(AUC_{ROC})$	$\sigma_{top}(AUC_{ROC})$	$\sigma^2_{top}(AUC_{ROC})$	$\sigma^2(AUC_{ROC})$	$\sigma^2_B(AUC_{ROC})$	$mean(RTR_{1\%})$	$mean(AUC_{ROC})$
Min. Div. Des. $k = 300$												
ACE inhibitors	0.40	0.0560	0.0023	0.0025	0.0002	0.89	0.0307	0.00094494	0.00097405	0.00002911	0.08	0.64
AChE Inhibitors	0.39	0.0682	0.0055	0.0057	0.0002	0.96	0.0096	0.00009121	0.00009310	0.00000188	0.11	0.80
Angio. R. Blockers	0.86	0.0330	0.0018	0.0019	0.0001	1.00	0.0005	0.00000029	0.00000031	0.00000001	0.19	0.86
COX inhibitors	0.19	0.0451	0.0020	0.0022	0.0001	0.91	0.0139	0.00019428	0.00020124	0.00000696	0.10	0.82
D2 antagonists	0.25	0.0567	0.0022	0.0024	0.0002	0.94	0.0094	0.00008752	0.00009108	0.00000356	0.13	0.85
HIV P. inhibitors	0.63	0.0926	0.0093	0.0095	0.0002	0.99	0.0061	0.00003711	0.00003750	0.00000039	0.16	0.78
5HT1A agonists	0.44	0.0542	0.0025	0.0027	0.0002	0.98	0.0038	0.00001433	0.00001479	0.00000046	0.14	0.84
5HT3 antagonists	0.70	0.0607	0.0028	0.0030	0.0002	0.99	0.0018	0.00000307	0.00000319	0.00000012	0.31	0.92
5HT reup. inhibitors	0.24	0.0511	0.0017	0.0019	0.0002	0.90	0.0152	0.00023023	0.00024106	0.00001084	0.16	0.84
PKC inhibitors	0.27	0.0778	0.0051	0.0053	0.0002	0.85	0.0242	0.00058382	0.00061249	0.00002868	0.10	0.67
Renin inhibitors	1.00	0.0005	0.0000	0.0000	0.0000	1.00	0.0002	0.00000003	0.00000003	0.00000000	0.74	0.98
Subst. P inhibitors	0.40	0.0643	0.0032	0.0034	0.0002	0.95	0.0106	0.00011248	0.00011598	0.00000350	0.17	0.88
Thrombin inhibitors	0.64	0.0795	0.0086	0.0088	0.0002	0.99	0.0053	0.00002843	0.00002876	0.00000033	0.20	0.81

Appendix B

Supplementary Tables: Chapter 3

Table B.1: Potential False Negatives. Overview.

AID	CID	# Refs. Found	Target	Mode of Interaction
429	661647	-	-	-
429	666976	-	-	-
429	2124942	-	-	-
429	2221829	-	-	-
429	3236183	-	-	-
449	797232	-	-	-
449	1219876	-	-	-
449	2140504	2	IRE1 endonuclease	inhibitor
449	3241605	-	-	-
552	650276	1	-	(decreased activity in motor activity test)
552	713182	-	-	-
552	762456	1	-	-
552	951589	-	-	-
552	3235718	-	-	-
524	645763	-	-	-
524	667157	-	-	-

Table B.1: Potential False Negatives. Overview.

AID	CID	# Refs. Found	Target	Mode of Interaction
524	870802	-	-	-
524	974684	-	-	-
524	2155052	1	p34cdc2 kinase	inhibition
525	644570	-	-	-
525	664796	-	-	-
525	2977461	-	-	-
525	6602980	-	-	-
565	664948	-	-	-
565	2945948	-	-	-
565	3236904	-	-	-
565	3240663	-	-	-
565	4094173	-	-	-
581	663452	-	-	-
581	741168	-	-	-
581	3237217	-	-	-
581	3238135	-	-	-
581	3244793	-	-	-
604	397588	2	-	-
604	870802	-	-	-
604	972417	-	-	-
604	2841612	1	-	-
604	5389712	-	-	-
628	653541	-	-	-
628	663038	-	-	-
628	1363188	-	-	-
628	2001334	-	-	-

Table B.1: Potential False Negatives. Overview.

AID	CID	# Refs. Found	Target	Mode of Interaction
628	2973815	-	-	-
629	654182	-	-	-
629	655077	-	-	-
629	725878	-	-	-
629	787302	-	-	-
629	3193162	1	stearoyl-CoA desaturase (SCD) enzymes (preferably SCD1)	inhibition
633	665223	-	-	-
633	795871	-	-	-
633	1247306	-	-	-
633	2632776	-	-	-
633	2646358	-	-	-
639	651906	-	-	-
639	795916	-	-	-
639	866943	-	-	-
639	2345962	-	-	-
639	3237613	-	-	-
641	291754	23	KDR, DANN-PK, RTK, CTK, STK	inhibition
641	646307	-	-	-
641	694544	1	-	-
641	717657	-	-	-
641	888381	-	-	-
689	403074	1	-	-
689	653588	-	-	-
689	1415442	-	-	-
689	2115030	-	-	-
689	2354012	-	-	-

Table B.1: Potential False Negatives. Overview.

AID	CID	# Refs. Found	Target	Mode of Interaction
727	286547	10	liver X receptors, p38, 5HT1-rec, vasopressin 1 rec.	activation/inhibition
727	657722	1	-	-
727	661489	1	E. coli primase	inhibition
727	807549	-	-	-
727	889075	-	-	-
736	658992	-	-	-
736	662422	-	-	-
736	1352540	-	-	-
736	2107624	-	-	-
736	2221133	2	-	-
798	951167	-	-	-
798	1280861	-	-	-
798	1444742	-	-	-
798	1444809	-	-	-
798	1718171	-	-	-
800	653266	-	-	-
800	714424	3	HIV-1 reverse transcriptase	inhibition
800	764249	-	-	-
800	2284606	-	-	-
800	2291974	-	-	-

Table B.2: Potential False Negatives. References.

AID	CID	References
449	2140504	[141,142]
552	650276	[143]
552	762456	[144]
524	2155052	[145]
604	397588	[146,147]
604	2841612	[148]
629	3193162	[149]
641	291754	[150–172]
641	694544	[173]
689	403074	[174]
727	286547	[175–184]
727	657722	[185]
727	661489	[186]
736	2221133	[187,188]
800	714424	[189–191]
800	764249	[192]

Table B.3: VS Figures of Merit for PubChem Datasets.

		Simple		MOE		SESP		MACCS	
Target	AID	mean	std	mean	std	mean	std	mean	std
AUC_{ROC}, 1Q., Orig.									
S1P1 rec.	466	0.551	0.053	0.620	0.078	0.544	0.035	0.568	0.075
PKA	548	0.718	0.106	0.711	0.125	0.678	0.074	0.744	0.096
SF1	600	0.564	0.056	0.538	0.028	0.580	0.048	0.545	0.049
Rho-Kinase2	644	0.615	0.090	0.581	0.095	0.601	0.045	0.627	0.073
HIV RT-RNase	652	0.596	0.084	0.513	0.068	0.500	0.098	0.550	0.071
Eph rec. A4	689	0.495	0.091	0.575	0.083	0.448	0.074	0.548	0.055
SF1	692	0.588	0.074	0.620	0.080	0.509	0.048	0.596	0.091
HSP 90	712	0.583	0.069	0.498	0.067	0.583	0.037	0.552	0.107
ER-a-Coact. Bind. Inh.	713	0.483	0.054	0.592	0.064	0.469	0.053	0.537	0.056
ER-ß-Coact. Bind. Inh.	733	0.519	0.067	0.523	0.067	0.471	0.053	0.534	0.073
ER-a-Coact. Bind. Pot.	737	0.553	0.095	0.680	0.064	0.552	0.083	0.616	0.086
FAK	810	0.510	0.060	0.518	0.064	0.587	0.064	0.597	0.085
Cathepsin G	832	0.638	0.052	0.638	0.080	0.613	0.061	0.736	0.066
FXIa	846	0.654	0.068	0.797	0.085	0.596	0.042	0.711	0.059
S1P2 rec.	851	0.484	0.153	0.522	0.072	0.443	0.105	0.494	0.121
FXIIa	852	0.646	0.057	0.766	0.065	0.667	0.077	0.769	0.063
D1 Rec.	858	0.504	0.026	0.524	0.052	0.451	0.027	0.502	0.066
M1 Rec.	859	0.560	0.076	0.583	0.084	0.580	0.047	0.589	0.113

Table B.3: VS Figures of Merit for PubChem Datasets.

		Simple		MOE		SESP		MACCS	
Target	AID	mean	std	mean	std	mean	std	mean	std
				AUC_{ROC}, 10Q., Orig.					
S1P1 rec.	466	0.584	0.033	0.699	0.022	0.564	0.015	0.649	0.033
PKA	548	0.798	0.034	0.861	0.031	0.745	0.029	0.844	0.021
SF1	600	0.593	0.032	0.642	0.042	0.633	0.025	0.657	0.048
Rho-Kinase2	644	0.729	0.040	0.770	0.054	0.679	0.037	0.802	0.039
HIV RT-RNase	652	0.650	0.055	0.592	0.029	0.465	0.087	0.638	0.036
Eph rec. A4	689	0.601	0.052	0.657	0.043	0.491	0.037	0.615	0.043
SF1	692	0.639	0.038	0.704	0.032	0.511	0.038	0.639	0.033
HSP 90	712	0.633	0.045	0.620	0.051	0.639	0.030	0.701	0.062
ER-a-Coact. Bind. Inh.	713	0.505	0.040	0.646	0.026	0.442	0.024	0.573	0.039
ER-ß-Coact. Bind. Inh.	733	0.529	0.036	0.560	0.034	0.453	0.023	0.572	0.036
ER-a-Coact. Bind. Pot.	737	0.605	0.048	0.782	0.039	0.538	0.040	0.698	0.043
FAK	810	0.576	0.031	0.600	0.035	0.627	0.033	0.687	0.039
Cathepsin G	832	0.724	0.035	0.785	0.049	0.693	0.030	0.880	0.045
FXIa	846	0.765	0.033	0.884	0.024	0.659	0.043	0.860	0.044
S1P2 rec.	851	0.627	0.082	0.676	0.075	0.577	0.088	0.702	0.093
FXIIa	852	0.759	0.023	0.845	0.018	0.753	0.022	0.879	0.029
D1 Rec.	858	0.513	0.021	0.556	0.025	0.456	0.015	0.538	0.038
M1 Rec.	859	0.615	0.026	0.655	0.039	0.631	0.012	0.691	0.034

Table B.3: VS Figures of Merit for PubChem Datasets.

		Simple		MOE		SESP		MACCS	
Target	AID	mean	std	mean	std	mean	std	mean	std
				$RTR_{1\%}$, 1Q., Orig.					
S1P1 rec.	466	0.020	0.019	0.025	0.020	0.021	0.019	0.034	0.037
PKA	548	0.086	0.088	0.132	0.140	0.092	0.086	0.123	0.109
SF1	600	0.016	0.019	0.029	0.029	0.025	0.026	0.032	0.036
Rho-Kinase2	644	0.053	0.065	0.071	0.072	0.067	0.070	0.089	0.067
HIV RT-RNase	652	0.031	0.028	0.037	0.046	0.043	0.035	0.042	0.050
Eph rec. A4	689	0.030	0.027	0.026	0.022	0.041	0.045	0.025	0.030
SF1	692	0.025	0.026	0.021	0.029	0.022	0.026	0.018	0.025
HSP 90	712	0.019	0.025	0.034	0.034	0.025	0.026	0.042	0.041
ER-a-Coact. Bind. Inh.	713	0.013	0.012	0.018	0.015	0.013	0.013	0.015	0.015
ER-ß-Coact. Bind. Inh.	733	0.012	0.013	0.015	0.015	0.017	0.014	0.021	0.015
ER-a-Coact. Bind. Pot.	737	0.033	0.033	0.056	0.043	0.027	0.020	0.028	0.024
FAK	810	0.023	0.021	0.026	0.023	0.036	0.027	0.036	0.027
Cathepsin G	832	0.042	0.031	0.070	0.045	0.063	0.055	0.131	0.084
FXIa	846	0.051	0.032	0.126	0.069	0.050	0.028	0.084	0.039
S1P2 rec.	851	0.067	0.076	0.069	0.074	0.065	0.073	0.074	0.069
FXIIa	852	0.072	0.072	0.122	0.094	0.094	0.088	0.192	0.137
D1 Rec.	858	0.011	0.008	0.013	0.012	0.013	0.009	0.014	0.015
M1 Rec.	859	0.022	0.023	0.029	0.025	0.031	0.033	0.035	0.036

Table B.3: VS Figures of Merit for PubChem Datasets.

		Simple		MOE		SESP		MACCS	
Target	AID	mean	std	mean	std	mean	std	mean	std
				$RTR_{1\%}$, 10Q., Orig.					
S1P1 rec.	466	0.041	0.019	0.059	0.024	0.059	0.027	0.093	0.038
PKA	548	0.215	0.061	0.198	0.066	0.301	0.080	0.281	0.081
SF1	600	0.047	0.024	0.097	0.041	0.069	0.033	0.129	0.048
Rho-Kinase2	644	0.080	0.036	0.101	0.038	0.135	0.062	0.208	0.057
HIV RT-RNase	652	0.054	0.027	0.112	0.042	0.044	0.040	0.133	0.046
Eph rec. A4	689	0.075	0.041	0.065	0.031	0.045	0.027	0.073	0.028
SF1	692	0.026	0.025	0.014	0.018	0.027	0.023	0.007	0.015
HSP 90	712	0.079	0.038	0.109	0.048	0.104	0.046	0.192	0.073
ER-a-Coact. Bind. Inh.	713	0.013	0.013	0.028	0.018	0.013	0.013	0.030	0.019
ER-ß-Coact. Bind. Inh.	733	0.010	0.010	0.031	0.018	0.015	0.010	0.065	0.024
ER-a-Coact. Bind. Pot.	737	0.023	0.021	0.112	0.045	0.037	0.026	0.042	0.034
FAK	810	0.059	0.027	0.063	0.024	0.068	0.027	0.089	0.036
Cathepsin G	832	0.111	0.037	0.189	0.054	0.218	0.065	0.404	0.083
FXIa	846	0.161	0.049	0.337	0.062	0.141	0.051	0.339	0.076
S1P2 rec.	851	0.231	0.097	0.285	0.114	0.238	0.097	0.323	0.123
FXIIa	852	0.193	0.055	0.301	0.056	0.218	0.061	0.436	0.062
D1 Rec.	858	0.013	0.011	0.022	0.013	0.021	0.012	0.024	0.013
M1 Rec.	859	0.049	0.023	0.046	0.021	0.094	0.027	0.061	0.030

Table B.3: VS Figures of Merit for PubChem Datasets.

		Simple		MOE		SESP		MACCS	
Target	AID	mean	std	mean	std	mean	std	mean	std
				AUC_{ROC}, 1Q., MUV					
S1P1 rec.	466	0.467	0.032	0.538	0.056	0.479	0.039	0.524	0.075
PKA	548	0.578	0.085	0.654	0.098	0.599	0.051	0.636	0.085
SF1	600	0.502	0.051	0.523	0.047	0.550	0.027	0.545	0.048
Rho-Kinase2	644	0.524	0.069	0.619	0.073	0.542	0.040	0.580	0.073
HIV RT-RNase	652	0.425	0.034	0.443	0.077	0.492	0.056	0.486	0.101
Eph rec. A4	689	0.514	0.093	0.549	0.089	0.457	0.095	0.544	0.065
SF1	692	0.503	0.052	0.563	0.067	0.462	0.037	0.544	0.083
HSP 90	712	0.473	0.059	0.485	0.095	0.517	0.049	0.536	0.120
ER-a-Coact. Bind. Inh.	713	0.425	0.038	0.512	0.085	0.452	0.033	0.491	0.068
ER-ß-Coact. Bind. Inh.	733	0.445	0.069	0.451	0.042	0.432	0.053	0.523	0.075
ER-a-Coact. Bind. Pot.	737	0.520	0.077	0.660	0.061	0.525	0.074	0.639	0.080
FAK	810	0.489	0.045	0.470	0.045	0.513	0.054	0.541	0.070
Cathepsin G	832	0.516	0.029	0.553	0.090	0.507	0.029	0.680	0.083
FXIa	846	0.515	0.059	0.692	0.116	0.492	0.040	0.616	0.064
S1P2 rec.	851	0.427	0.081	0.467	0.049	0.451	0.056	0.471	0.085
FXIIa	852	0.499	0.028	0.647	0.104	0.530	0.068	0.687	0.104
D1 Rec.	858	0.458	0.026	0.492	0.065	0.488	0.029	0.461	0.070
M1 Rec.	859	0.438	0.033	0.513	0.057	0.525	0.034	0.518	0.067

Table B.3: VS Figures of Merit for PubChem Datasets.

		Simple		MOE		SESP		MACCS	
Target	AID	mean	std	mean	std	mean	std	mean	std
				AUC_{ROC}, 10Q., MUV					
S1P1 rec.	466	0.462	0.054	0.553	0.050	0.525	0.050	0.580	0.061
PKA	548	0.571	0.059	0.732	0.045	0.626	0.041	0.718	0.039
SF1	600	0.494	0.056	0.600	0.051	0.582	0.038	0.594	0.052
Rho-Kinase2	644	0.528	0.048	0.696	0.043	0.601	0.047	0.743	0.056
HIV RT-RNase	652	0.458	0.060	0.502	0.054	0.498	0.047	0.530	0.066
Eph rec. A4	689	0.535	0.065	0.627	0.054	0.460	0.069	0.617	0.057
SF1	692	0.485	0.045	0.623	0.040	0.453	0.041	0.592	0.043
HSP 90	712	0.497	0.065	0.570	0.060	0.568	0.037	0.658	0.074
ER-a-Coact. Bind. Inh.	713	0.458	0.058	0.611	0.047	0.496	0.045	0.502	0.046
ER-ß-Coact. Bind. Inh.	733	0.482	0.045	0.510	0.049	0.445	0.042	0.565	0.041
ER-a-Coact. Bind. Pot.	737	0.510	0.054	0.739	0.044	0.503	0.052	0.709	0.047
FAK	810	0.524	0.042	0.524	0.048	0.543	0.050	0.672	0.039
Cathepsin G	832	0.535	0.054	0.711	0.066	0.554	0.048	0.864	0.059
FXIa	846	0.524	0.038	0.824	0.041	0.533	0.057	0.793	0.054
S1P2 rec.	851	0.465	0.065	0.579	0.076	0.589	0.072	0.660	0.092
FXIIa	852	0.522	0.044	0.739	0.038	0.620	0.044	0.796	0.038
D1 Rec.	858	0.490	0.055	0.605	0.055	0.521	0.051	0.546	0.066
M1 Rec.	859	0.467	0.057	0.550	0.053	0.527	0.038	0.603	0.053

Table B.3: VS Figures of Merit for PubChem Datasets.

		Simple		MOE		SESP		MACCS	
Target	AID	mean	std	mean	std	mean	std	mean	std
$RTR_{1\%}$, 1Q, MUV									
S1P1 rec.	466	0.012	0.016	0.020	0.026	0.024	0.029	0.020	0.026
PKA	548	0.015	0.024	0.043	0.060	0.029	0.032	0.036	0.040
SF1	600	0.012	0.016	0.022	0.030	0.021	0.036	0.029	0.047
Rho-Kinase2	644	0.011	0.018	0.036	0.031	0.033	0.033	0.038	0.038
HIV RT-RNase	652	0.016	0.021	0.013	0.024	0.023	0.023	0.018	0.022
Eph rec. A4	689	0.013	0.020	0.021	0.028	0.031	0.038	0.020	0.031
SF1	692	0.008	0.014	0.012	0.017	0.007	0.016	0.009	0.019
HSP 90	712	0.011	0.021	0.027	0.040	0.014	0.018	0.029	0.036
ER-a-Coact. Bind. Inh.	713	0.009	0.015	0.020	0.024	0.012	0.018	0.017	0.021
ER-ß-Coact. Bind. Inh.	733	0.006	0.016	0.015	0.024	0.018	0.027	0.021	0.025
ER-a-Coact. Bind. Pot.	737	0.014	0.018	0.033	0.032	0.019	0.019	0.024	0.027
FAK	810	0.006	0.017	0.013	0.017	0.014	0.022	0.027	0.029
Cathepsin G	832	0.010	0.021	0.048	0.044	0.027	0.038	0.081	0.055
FXIa	846	0.003	0.009	0.068	0.061	0.023	0.022	0.040	0.033
S1P2 rec.	851	0.033	0.050	0.044	0.056	0.043	0.056	0.052	0.052
FXIIa	852	0.008	0.015	0.040	0.054	0.027	0.033	0.073	0.075
D1 Rec.	858	0.014	0.020	0.022	0.024	0.017	0.022	0.018	0.024
M1 Rec.	859	0.012	0.018	0.018	0.026	0.014	0.022	0.014	0.024

Table B.3: VS Figures of Merit for PubChem Datasets.

		Simple		MOE		SESP		MACCS	
Target	AID	mean	std	mean	std	mean	std	mean	std
				$RTR_{1\%}$, 10Q., MUV					
S1P1 rec.	466	0.001	0.007	0.042	0.032	0.042	0.038	0.077	0.047
PKA	548	0.000	0.000	0.042	0.035	0.097	0.057	0.047	0.040
SF1	600	0.000	0.000	0.050	0.038	0.026	0.029	0.063	0.051
Rho-Kinase2	644	0.000	0.000	0.079	0.050	0.079	0.049	0.083	0.037
HIV RT-RNase	652	0.000	0.000	0.026	0.028	0.054	0.039	0.041	0.032
Eph rec. A4	689	0.001	0.005	0.027	0.025	0.002	0.009	0.048	0.035
SF1	692	0.000	0.000	0.006	0.016	0.000	0.000	0.001	0.007
HSP 90	712	0.000	0.000	0.043	0.037	0.036	0.030	0.051	0.048
ER-a-Coact. Bind. Inh.	713	0.000	0.000	0.077	0.043	0.012	0.025	0.064	0.041
ER-ß-Coact. Bind. Inh.	733	0.000	0.000	0.021	0.025	0.026	0.025	0.057	0.052
ER-a-Coact. Bind. Pot.	737	0.000	0.000	0.025	0.026	0.018	0.025	0.014	0.022
FAK	810	0.000	0.000	0.004	0.015	0.024	0.025	0.074	0.047
Cathepsin G	832	0.000	0.000	0.126	0.053	0.085	0.048	0.331	0.095
FXIa	846	0.000	0.000	0.230	0.086	0.010	0.021	0.227	0.089
S1P2 rec.	851	0.101	0.055	0.253	0.093	0.185	0.072	0.306	0.116
FXIIa	852	0.000	0.000	0.119	0.062	0.028	0.036	0.237	0.071
D1 Rec.	858	0.000	0.000	0.084	0.049	0.029	0.036	0.087	0.052
M1 Rec.	859	0.000	0.000	0.017	0.024	0.052	0.035	0.044	0.030

Appendix C

MUV Datasets

The MUV datasets are completely based on structural information and bioactivity data from PubChem. This data is freely available, but it is not allowed to re-distribute the structures in SDF format. The MUV datasets can be automatically downloaded using the MUV_Downloader, a JAVA program that queries the PubChem Power User Gateway (PUG) for the structures and downloads the datasets in .sdf.gz format to a specified directory on your computer.

The MUV downloader is available on the enclosed CD-ROM in the directory

```
/MUV_Downloader
```

or via the internet at

```
http://www.pharmchem.tu-bs.de/lehre/baumann/download.html
```

A Sun JAVA SE6 compatible JRE must be installed on your system for the MUV_Downloader to work. It can be downloaded free of charge from Sun:

```
http://www.java.com/en/download/manual.jsp
```

Once you have downloaded MUV_Downloader.jar and JAVA is installed correctly on your system, you can start the program by typing

```
~/$ java -jar MUV_Downloader.jar
```

at your system's command prompt. On most systems you should also be able to run the MUV_Downloader by simply double clicking its symbol in the file manager.

The complete collection of MUV datasets contains 255510 compounds (17 * 30 actives + 17*15000 decoys), so the download will take quite a while.

Appendix D

Spatial Statistics Toolbox for MATLAB 7

D.1 Installation

The implementation of the algorithms for the spatial statistics analysis of chemical datasets constituted a major effort in the completion of this study. The programs for the calculation of $G(t)$, $F(t)$ and the resulting scalars ΣG, ΣF and ΣS, as well as those for the design of datasets with given topological properties are the core of a Toolbox for The Mathworks MATLAB 7[90] that is provided on the enclosed CD-ROM. The source code of the most important programs will be listed in print in the following sections. (Sections D.2.1, D.2.2, D.2.3 and D.2.4)

In addition to the algorithms and programs utilized and discussed in this study, the toolbox provides a number of programs for the calculation of other established statistics for the analysis of mapped point patterns. For a comprehensive review of these methodologies refer to Ref.[94] Moreover, the toolbox features a JAVA helper library for the handling of chemical structures and the calculation of molecular descriptors, which is a prerequisite for the conduction of spatial statistics analyses of chemical data. Using the Spatial Statistics Toolbox and the JAVA helper library, users can apply the methods introduced in this study to their own datasets. For example, the toolbox can be used to

design custom MUV datasets from user-provided datasets. Section D.3 features a HowTo tutorial demonstrating the generation of MUV datasets based on two example datasets.

In order to install the Spatial Statistics Toolbox copy the folder

```
/SpSt
```

from the root directory of the CD-ROM to your hard drive. Alternatively, you can download the toolbox as a .tar.gz archive from

```
http://www.pharmchem.tu-bs.de/lehre/baumann/SpaceStatsToolbox.html
```

In this case please extract the contents of the archive to a location of your choice. Finally, include the folder

```
SpSt/
```

on your disk *recursively* into your MATLAB path. You can do so by clicking File->Set Path...->Include with subfolders... in the MATLAB workspace.

As mentioned before, the toolbox is complemented by a JAVA library jMUV.jar managing SD-file I/O and descriptor calculation. jMUV itself is based on the Chemistry Development Kit CDK. (Version 1.0.3, Download) Both libraries have to be included in your MATLAB classpath. They are available in the directory

```
/jar
```

on the enclosed CD-ROM or can be downloaded from

```
http://www.pharmchem.tu-bs.de/lehre/baumann/jMUV.jar
http://downloads.sourceforge.net/cdk/cdk-1.0.3.jar
```

respectively. Both libraries have to be included in the MATLAB classpath. Please be aware, that jMUV.jar is built based on Version 1.0.3 of the CDK. Other versions might not work properly.

D.2 Source Code

D.2.1 $G(t)$ - spst_G

```
function G = spst_G(D, map, options)
%
% G = spst_G(D, map, options)
%
% Input:    D       Data matrix with rows = observations, columns=variables
%
%           map     Hypercubic map for uniform csr. For higher data
%                   dimensions (>3) the use of a map and corresponding edge
%                   correction is strongly advised against. Use map = [] instead.
%
%           options Options struct variable. Default values are indicated
%                   by *asterisks*.
%
%           options.distmode:   *'euc'*, 'city', 'cheby'
%           options.step:       *0.01*, any decimal
%           options.maxD:       *10*, any integer
%
% Copyright:        Sebastian Rohrer
%                   University of Braunschweig, Institute of Technology
%                   Department of Pharmaceutical Chemistry
%                   2008
[m n] = size(D);
% if options is omitted, set default values
if (nargin < 3)
    options.distmode='euc';
    options.step=0.01;
    options.maxD=10;
end
% set x-axis and preallocate G
x = 0:options.step:options.maxD;
G = zeros(size(x,2), 1);
% if a map is used calculate distance to border for edge correction
if (max(size(map))>0)
    d2b_events = dist2border(D, map);
    d2b_events = d2b_events * ones(1,size(x,2));
end
% For each event calculate distance to nearest neighbor
dist = nndist(D, D, 1, options.distmode);
dist = dist*ones(1,size(x,2));
% calculate cumulative distribution
w = ones(m,1)*x;
dnn = dist < w;
% if a map is used, apply edge correction
if (map)
    in_center = d2b_events <= w;
    nn = dnn & in_center;
else
    nn = dnn;
end
G = (sum(nn)./m)';   % We want the fraction of events
G = [x' G];  % Append x-axis
```

D.2.2 $F(t)$ - spst_F

```
function F = spst_F(D, map, I, options)
% F = spst_F(D, map, I, options)
%
% Input:    D       Data matrix with rows = observations, columns=variables
%
%           map     Hypercubic map for uniform csr. For higher data
%                   dimensions (>3) the use of a map and corresponding uniform csr
%                   is strongly advised against. Use map = [] and
%                   options.csr = 'bt' instead.
%
%           I       Background data for the generation of bootstrap or
%                   convex pseudo-data csr.
%
%           options Options struct variable. Default values are indicated
%                   by *asterisks*.
%
%           options.distmode:   *'euc'*, 'city', 'cheby'
%           options.csr:        *'bt'*, 'pseudo', 'disc', 'dec', 'all'
%           options.nP:         *10000*, any integer
%           options.iter:       *20*, any integer
%           options.step:       *0.01*, any decimal
%           options.maxD:       *10*, any integer
%
% Copyright:        Sebastian Rohrer
%                   University of Braunschweig, Institute of Technology
%                   Department of Pharmaceutical Chemistry
%                   2008
%initialize random numbers algorithm
old = rand('state');
rand('state', 0);
% if options is omitted, set default values
if (nargin < 4)
    options.distmode='euc';
    options.csr='bt';
    options.nP=10000;
    options.iter=1;
    options.step=0.01;
    options.maxD=10;
end

% set x-axis and preallocate F
x = 0:options.step:options.maxD;
F = zeros(size(x,2), options.iter);
% start calculation
for j=1:options.iter
    nP = options.nP;
    % generate random points
    if strcmpi(options.csr, 'bt')
        sP = bootstrap(nP, I,options.replacement);

    elseif strcmpi(options.csr, 'all')
        sP = I;
        nP = size(sP,1);
    elseif strcmpi(options.csr, 'pseudo')
        sP = convex_pseudo_data(nP, I);
    elseif strcmpi(options.csr, 'dec')
        sP = unifcsr(map, nP, 'dec');
    elseif strcmpi(options.csr, 'disc')
```

```
        sP = unifcsr(map, nP, 'disc');
    else
        error('spst_F: CSR (options.csr) mode not supported.');
    end

    % if a map is used calculate distance of each point to all events
    % else compute only distance to nearest event
    if (max(size(map))>0)
        dist = distance(sP, D, options.distmode);
        % calculate distance to border for edge correction
        d2b_events = dist2border(D, map);
        d2b_points = dist2border(sP, map);
    else
        dist = nndist(sP, D, 1, options.distmode);
    end

    % calculate cumulative distribution
    for i=1:size(x,2);
        w = x(i);

        % if a map is used, apply edge correction
        if (map)
            events = d2b_events <= w;
            points = d2b_points <= w;
            dnn = min(dist(points,events), [], 2);
        else
            dnn = dist;
        end

        nn = dnn < w;

        num_set = sum(nn);
        F(i,j) = num_set;
    end
end
F= F./nP;       % We want the fraction of points
F=mean(F,2);    % Get the mean from the iterations
F = [x' F];     % Append x-axis
rand('state', old); % Return random number generator to its original state.
```

D.2.3 Design of Datasets of Actives with Given ΣG - spst_RX

```
function [R, sgs] = spst_RX(A, Red, nsel, options)
%
% [R, sff] = spst_RX(A, Red, nsel, options)
%
% Row-exchange algorithm to optimize a sample of actives towards a preset value
% of SigmaG
%
%
% Input:    A       Data matrix of actives with rows = compounds,
%                   columns=descriptors
%
%           Red     Logical array constituting the starting design of
%                   actives generated by spst_ksnn
%
%           nsel    Number of actives to select
%
```

```
%           options Options struct variable. Default values are indicated
%                   by *asterisks*.
%
%           options.distmode:   *'euc'*, 'city', 'cheby'
%           options.iter:       *20*, any integer
%           options.step:       *0.01*, any positive number
%           options.maxD:       *10*, any positive number
%           options.targetG:    *312*, any positive number
%           options.deltaSigmaG *2*, any positive number
%           options.r           *0.8*, any positive number
%           options.verbose     *true*, boolean
%
% Copyright:        Sebastian Rohrer
%                   University of Braunschweig, Institute of Technology
%                   Department of Pharmaceutical Chemistry
%                   2008
% initialize random numbers generator
old = rand('state');
rand('state', 0);
% get dataset size
[m n] = size(A);
% initialize return values
R = false(m,1);
sgs = zeros(options.iter,2);
% initialize design variables
d = zeros(nsel,1);
sG = options.targetG;
% Starting design from previous step (Kennard-Stone)
d(:,1) = find(Red);
% get stopping criterion
threshold=options.deltaSigmaG;
% initiate tabu matrix
% 0 means tabu for next iteration
tabu = ones(m,1);
tabu(d,1) = 0;
% initialize loop variables and off we go...
iter = 1;
has_changed = 1;
is_optimized = 0;
while(iter <= options.iter & has_changed == 1 & is_optimized == 0)
    has_changed = 0;

    % loop over all actives in the currently selected set
    for i=1:nsel

        % compute of the set d(SigmaG) before exchange
        G = spst_G(A(d,:), [], options);
        G = sum(G(:,2));
        old_DsG = abs(G - sG);

        % throw out row i without using setdiff (terribly slow)
        d2 = [d(1:(i-1),1);d((i+1):end)];

        % find rows not yet in set
        nd = find(tabu);

        % calculate nearest neighbor distances
        nndistance = nndist(A(nd,:), A(d,:), 1, options.distmode);

        % remove candidates below similarity cutoff and outliers
```

```
        nd = nd(nndistance >= options.r);

        if(nd)  % proceed only if there are selectable candidates left

            % compute sG for exchange
            sgg = zeros(size(nd,1),1);
            for j=1:size(nd,1)
                gg = spst_G(A([d2;nd(j,1)],:), [], options);
                sgg(j,1) = sum(gg(:,2));
            end
            % the row that produces the lowest difference from sG is the best for exchange
            dsgg = abs(sG-sgg);
            % Now we want the point creates the smallest differences from
            % sG
            [mdsgg, ind] = min(dsgg);
            new_DsG = dsgg(ind,1);
            new_sG = sgg(ind,1);
            new_member = nd(ind,1);
            % make exchange only if DsG decreases
            if (old_DsG - new_DsG > 0)
                tabu(d(i,1),1) = 1;             % allow replaced candidate to replace the other rows
                d(i,1) = new_member;            % replace old candidate by new one
                tabu(new_member,1) = 0;         % disallow new_member for following replacements
                sgs(iter,1) = new_DsG;          % record old and new DsG
                sgs(iter,2) = new_sG;
                has_changed = 1;                % indicate that the current iteration has changed the set

                if(new_DsG <= threshold)        % if optimization criterion is reached exit the loop
                    is_optimized = 1;
                    break;
                end
            else                                % if the exchange is not
                                                % favorable (and therefore not made)...
                sgs(iter,1) = old_DsG;          % record DsG
                sgs(iter,2) = G;
            end
        else
            sgs(iter,1) = old_DsG;         % record DsG
            sgs(iter,2) = G;
        end

    end
    iter = iter + 1;                       % increment interation counter
end
if (options.verbose)
    disp(strcat('Optimized after:', num2str(iter), ' Iterations. delta(SigmaG)='
      , num2str(sgs(iter-1,1)), ' SigmaG=', num2str(sgs(iter-1,2))));
end
if(iter == options.iter && options.verbose)
    disp('Maximum Iterations reached');
end
d = sort(d,'ascend');
R(d,1) = true;
rand('state', old);
```

D.2.4 Design of Datasets of Decoys with Given ΣF - spst_GA

```
function [R, sff] = spst_GA(A, I, Red, nsel, options)
```

```
%
% [R, sff] = spst_GA(A, I, Red, nsel, options)
%
% Genetic algorithm to optimize a sample of decoys towards a preset value
% of SigmaF
%
%
% Input:    A       Data matrix of actives with rows = compounds,
%                   columns=descriptors
%
%           I       Data matrix of decoys
%
%           Red     Logical array constituting the starting design of
%                   decoys generated by spst_ksnn
%
%           nsel    Number of decoys to select
%
%           options Options struct variable. Default values are indicated
%                   by *asterisks*.
%
%           options.distmode:   *'euc'*, 'city', 'cheby'
%           options.iter:       *20*, any integer
%           options.step:       *0.01*, any decimal
%           options.maxD:       *10*, any integer
%           options.targetG:    *312*, any positive number
%           options.deltaSigmaG *2*, any positive number
%           options.verbose     *true*, boolean
%
% Copyright:        Sebastian Rohrer
%                   University of Braunschweig, Institute of Technology
%                   Department of Pharmaceutical Chemistry
%                   2008
% initialize random numbers generator
% This will generate reproducible results
% If you want true randomness, comment out the following two and the last
% line (reset rand) of the function.
old = rand('state');
rand('state', 0);
% get fitness threshold from options
threshold = options.deltaSigmaG;
% initialize iteration counter
iter = 0;
% set optimization state
optimized = false;
% initialize return values
R = false(size(I,1),1);
sff = zeros(options.iter,1);
% initialize the genepool (all potential decoys)
gp = true(size(I,1),1);
if(options.verbose)
    disp('Calculating NN-Distances...');
end
% remove candidates too far off (farther than maximum NN-Dist in A);
annd = max(nndist(A, A, 1, options.distmode));
% Innd (precalculated NN distances of Inactives) is also used later for
% fitness calculation see function sigmaF = sigmaF_precompNND (below)
Innd = nndist(I, A, 1, options.distmode);
gp(Innd > annd,1) = false;
% convert logical to indices
gpi = find(gp);
```

```
% exit with error, if number of candidates too small
if(size(gpi,1) <= nsel)
    error('spst_Fga: number of candidates too small!');
end
% get reasonable parent from kNN design prduced in previous step
PRI = find(Red);
% generate npop starting individuals by mutating the parent
P = [PRI spst_ga_mutate(PRI, gpi, 0.3, options.npop-1)];
% compute fitness of starting population
% (the smaller, the better!)
if(options.verbose)
    disp('Calculating first fitness vector...');
end
f = spst_Fga_fitness(Innd, P, options);
% sort ascending
[f, ind] = sort(f);
P = P(:,ind);
bestF = f(1);
if(options.verbose)
    % output fitness of the five best individuals
    disp(strcat('Fitness of starting design:', num2str(f(1:5))));
end
if (bestF <= threshold)
        optimized = true;
end
if (options.verbose)
    disp('Start evolution...');
end
no_change_count=0;
while(iter < options.iter && not(optimized) && no_change_count <=3)



    % initialize children
    P2 = zeros(size(P));

    % keep the 3 best individuals for the next generation
    P2(:,1:3) = P(:,1:3);

    % generate 10 children by mutating best five with low mutation rate
    P2(:,4:13) = spst_ga_mutate(P(:,1:5), gpi, 0.1, 10);

    % generate 10 children by mutating best five with high mutation rate
    P2(:,14:23) = spst_ga_mutate(P(:,1:5),gpi, 0.3, 10);

    % generate 60 children with very five mutation rate
    P2(:,24:83) = spst_ga_mutate(P(:,1:5),gpi, 0.05, 60);

    % generate remaining children by crossover of best 10
    P2(:,84:end)= spst_ga_crossover(P(:,1:10), options.npop-83);

    % add a slight level of mutation to a third of the crossover-children
    h = floor(size(P2(:,84:end),2)/3);
    P2(:,84:(83+h)) = spst_ga_mutate(P2(:,84:(83+h)),gpi, 0.02, h);

    % fitness of new population (the smaller, the better!)
    f2 = spst_Fga_fitness(Innd, P2, options);
    % sort ascending
    [f2, ind] = sort(f2);
    P2 = P2(:,ind);
```

```
    % get best fitness
    bestF2 = f2(1);
    % log best fitness value for later analysis
    sff(iter+1,1) = bestF2;
    % set new population
    P = P2;

    % exit loop if optimization threshold is reached;
    if (bestF2 < bestF)
        bestF = bestF2;
        no_change_count=0;
    else
        no_change_count=no_change_count+1;
    end

    if (options.verbose)
        % output fitness of the five best individuals
        disp(strcat(num2str(iter+1), ' Iterations. Fitness:', num2str(f2(1:5)),
         ', NoChangeCount:', num2str(no_change_count)));
    end

    if (bestF <= threshold)
        optimized = true;
    end
    iter = iter+1;
end
if (options.verbose)
    % output optimization result
    disp(strcat('spst_Fga converged after:', num2str(iter),
      ' Iterations. With a final fitness of:', num2str(bestF)));
end
R(P(:,1),1) = true;
% reset rand
rand('state', old);
function f = spst_Fga_fitness(Innd, P, options)
%
% This function calculates SigmaF for all individuals of a population and
% compares it to the target value. The difference between SigmaF of each
% individual and the target value (options.targetG) is the fitness of the
% individual.
% initialize
sf = zeros(1, size(P,2));
% loop through individuals and compute sigmaF (see below)
for i=1:size(P,2)
    sf(1,i) = sigmaF_precompNND(Innd(P(:,i),:), options);
end
% fitness = absolute difference
f = abs(options.targetG - sf);
function sigmaF = sigmaF_precompNND(nndist, options)
%
% By using a precomputed vector of the distance of each decoy to its
% nearest neighbor active, this function calculates F and sigmaF for a
% sample of decoys much faster than spst_F.
% determine dataset sizes
[m n] = size(nndist);
% Initialize x-Axis
x = ones(m,1)*(0:options.step:options.maxD);
% Inflate the distance vector to a matrix the same size as x
d = nndist*ones(1,size(x,2));
```

```
% Calculate F
nn = d < x;
num_set = sum(nn);
F= num_set./m;
% Calculate and return SigmaF
sigmaF = sum(F);
function C = spst_ga_mutate(P, gpi, mutRate, numC)
%
% This function mutates a sample of decoys (an individual) by replacing
% part of it by random decoys selected from the genepool (all other
% potential decoys).
% determine population size
[m n] = size(P);
% if we want to produce more children than parents
if (numC>n)
    % repeat P until there are more than enough parents
    size_factor = ceil(numC/n);
    P = repmat(P,1,size_factor);
    % randomly select enough parents
    r = randperm(size_factor*n);
    P=P(:,r(1:numC));
end
% if we want to produce less children than parents
if(numC<n)
    r = randperm(n);
    P = P(:,r(1:numC));
end
C = zeros(m,numC);
for i=1:numC
    % disallow decoys already present in the current set
    gpii = setdiff(gpi, P(:,i));

    % mutate according to mutation rate
    num_keep = floor(m*(1-mutRate));
    num_get = m-num_keep;

    % randomly select the decoys to keep
    r1 = randperm(m);
    k = P(r1(1:num_keep),i);
    % fill up other places with randomly chosen decoys from the pool.
    r2 = randperm(size(gpii,1));
    g = gpii(r2(1:num_get));

    C(:,i) = [k; g];
end
function C = spst_ga_crossover(P, numC)
%
% This function generates a new sample of decoys (child chromosome) by
% randomly combining two other samples of decoys (parent chromosomes).
[m n] = size(P);
C = zeros(m,numC);
% if we want to produce more children than parents
if (numC>n)
    % repeat P until there are more than enough parents
    size_factor = ceil(numC/n);
    P = repmat(P,1,size_factor);
    % randomly select enough parents
    r = randperm(size_factor*n);
    P=P(:,r(1:numC));
end
```

```
% generate random pairs for crossover.
p = zeros(2,numC);
p(1,:) = randperm(numC);
p(2,:) = randperm(numC);
% do actual crossover
for j=1:numC
    % combine both parents
    sex = unique([P(:,p(1,j));P(:,p(2,j))]);
    % mix randomly
    r = randperm(size(sex,1));
    sex = sex(r);

    % create the child
    C(:,j) = sex(1:m,1);
end
```

D.3 How To Generate Your Own MUV Datasets

D.3.1 Basic Preparations

This tutorial relies on two example datasets. These datasets are available on the enclosed CD-ROM in the directory

```
/example_data
```

or as a .tar.gz archive from

```
http://www.pharmchem.tu-bs.de/lehre/baumann/examples.tar.gz
```

Please copy the files to a folder on your hard drive. We will refer to the respective folder as

```
'example_dir'
```

from now on.

The MUV workflow is designed to generate a collection of VS benchmarking datasets from a collection of datasets with bioactivity against several biological targets. We will refer to these datasets as *activity classes.* We will refer to the name of such an activity class by the variable

```
classe
```

(class is a MATLAB keyword and can not be used as a variable name) from now on. Each activity class consists of one SD-file of compounds that are known to be *active* against the respective target and another SD-file of compounds that are known or assumed to be *inactive* against the target. Since the MUV workflow selects corresponding subsets of actives and inactives to generate spatially optimal active and decoy datasets, we will refer to the initial files as *potential actives (PA)* and *potential decoys (PD)* respectively. For MUV to work, all files of potential actives and potential decoys have to be collected in a common directory and be named following certain conventions. Each corresponding pair of potential actives and decoys must be named like:

```
classe_suffix.sdf
```

with classe the class name (e.g. ACE or Renin, Bioassay456, ...) and suffix identifying actives and inactives (e.g. suffix='actives.sdf')

The files in the 'example_dir' directory adhere to these conventions:

```
classe1_actives.sdf   -  Example dataset classe1, actives
classe1_decoys.sdf    -  Example dataset classe1, decoys
classe2_actives.sdf   -  Example dataset classe2, actives
classe2_decoys.sdf    -  Example dataset classe1, decoys
```

Before you start, you have to provide MATLAB with a cell-array of the class names and the suffixes. For the example datasets type

```
>> classes = {'classe1'; 'classe2'};
>> suffix = {'actives'; 'decoys'};
```

at the MATLAB prompt.

The spatial statistics toolbox manages options that are needed in multiple function using a MATLAB struct variable called options. In order to obtain an options struct with default values, type

```
>> options=spst_getDefaults
```

at the MATLAB prompt, which will produce the following output:

```
options =
       distmode: 'euc'
       csr: 'bt'
       nP: 10000
       iter: 20
       step: 0.0100
       maxD: 10
       weight: 'fortin'
       normalize: 1
       sort: 0
       replacement: 1
       r: 0.8000
       npop: 150
       deltaSigmaG: 2
       verbose: 1
       targetG: 312
       sep: '/'
```

So far, the only important options are .normalize which should be left in its default state of 'true' and .sep which constitutes the directory separator on your system. On Unix systems it can be left '/' on Windows systems, it should be changed to '\'.

D.3.2 Calculate Simple Descriptors

As mentioned above, Maximum Unbiased Validation is based on a spatial optimization of datasets in simple descriptor space. Therefore you have to calculate these descriptors for your datasets first. There are two ways of doing so: (i) If you are lucky enough to have a valid license for FILTER and BABEL by OpenEye Inc., MUV provides a wrapper that utilizes these tools for simple descriptor calculation. (ii) If you don't have access to OpenEye's tools, MUV also provides an implementation of simple descriptors based on the open source Chemistry Development Kit (CDK).

At the present state, simple descriptor calculation is much faster and reliable using the OpenEye tools. Apparently, there are problems in the CDK with the correct detection of chiral centers and the correct prediction of the logP, which are both important variables in

simple descriptors. The respective bug reports have been filed to the CDK bug tracking system.

The function muv_simple_descriptors provides functionality for the calculation of simple descriptors using both, OpenEye tools and the CDK.

D.3.2.1 Calculating Simple Descriptors using OpenEye FILTER and BABEL

As a first measure you need to provide the function with your system's commands to FILTER and BABEL. These are hard-coded in lines 43 and 44 of the function. Before first use, you must edit these two lines to match your system. On my system, FILTER and BABEL are linked to /usr/bin and therefore the respective lines in muv_simple_descriptors simply are:

```
filtercmd='filter';
(or filtercmd='/opt/OpenEye/bin/filter...)
babelcmd='babel';
(or babelcmd='C:\OpenEye\bin\babel', ... )
```

Furthermore, you have to provide a scratch directory with read and write access. This is necessary for saving the text output of FILTER and BABEL, which is parsed to yield the descriptors. The location of the scratch directory is specified in line 47 of the function, on my system it is:

```
scratch='/home/baschti/scratch';
(Don't use abbreviations like '~/scratch')
```

The numerical values of ranges of the variables in simple descriptors are different. Therefore the descriptor matrices should be autoscaled columnwise. In order to do so, the mean and standard deviation for each simple descriptor variable have been determined for a very large and comprehensive collection of compounds. The respective values are hard-coded in muv_simple_descriptors and can be readily used for normalizing your data. In order to do so set

```
>>options.normalize=true;
```

which is also the default. Now you're good to go. At the MATLAB prompt type

```
>>act = muv_simple_descriptors('example_dir', classes, suffix{1}, 'OE', options)
Output:
This program calculates simple descriptors for a directory of SD-files.
The content of the descriptor vectors is:
#B #Br #C #Cl #F #I...
Calucalting simple descriptors for:example_data/classe1_actives.sdf...
Calculating Properties.
Calculating AtomCounts.
...done.
Scaling data...
Calucalting simple descriptors for:example_data/sdf/classe2_actives.sdf...
Calculating Properties.
Calculating AtomCounts.
...done.
Scaling data...
act =
    classe1: [1x1 struct]
    classe2: [1x1 struct]
>>dec = muv_simple_descriptors('example_dir', classes, suffix{2}, 'OE', options)
Output:
This program calculates simple descriptors for a directory of SD-files.
The content of the descriptor vectors is:
#B #Br #C #Cl #F #I...
Calucalting simple descriptors for:example_data/classe1_decoys.sdf...
Calculating Properties.
Calculating AtomCounts.
...done. Scaling data...
Calucalting simple descriptors for:example_data/classe2_decoys.sdf...
Calculating Properties.
Calculating AtomCounts.
...done.
Scaling data...
dec =
    classe1: [1x1 struct]
    classe2: [1x1 struct]
```

What you did was calling muv_simple_desc first for the actives (suffix{1}) and then for the decoys (suffix{2}). This will provide you with two struct variables act and dec containing a field for each activity class, each with a subfield .dsc containing the actual descriptors and .ids the compound IDs as given by the first line in each molecule block in the SD-files. The descriptor calculation should be done in a moment for the actives (~500 compounds), but might take quite a while for the decoys (~130000 compounds) depending on your hardware. First benchmarks indicate a processing time of roughly 0.01 sec per compound on an Intel Core2Duo (1.66 GHz, 4GB RAM, Ubuntu Linux 7.10).

D.3.2.2 Calculating Simple Descriptors Using the CDK

With jMUV.jar and cdk-1.0.3.jar in your MATLAB classpath, there are no further preparations necessary for the calculation of simple descriptors using the CDK. Set

```
>>options.normalize=true;
```

and call

```
>>act = muv_simple_descriptors('example_dir', classes, suffix{1}, 'CDK', options)
>>dec = muv_simple_descriptors('example_dir', classes, suffix{2}, 'CDK', options)
```

As with the OpenEye tools, this will provide you with two struct variables act and dec containing a field for each activity class, each with a subfield .dsc containing the actual descriptors and .ids the compound IDs as given by the first line in each molecule block in the SD-files. As stated above, we discourage the use of the CDK for simple descriptor calculation at the present state. In addition to the lower quality of the descriptors calculated by the CDK, descriptor calculation is about 10x slower than with the OpenEye Tools.

D.3.3 Generating Maximally Diverse Sets of Actives and Minimally Separated Sets of Decoys

The goal of MUV design is to adjust all benchmark datasets in a collection to a common degree of clumping (ΣS) in simple descriptor space. The state of spatial randomness $\Sigma S = 0$ has proven to be especially advantageous. In the MUV workflow, all datasets of actives are first adjusted to a common level of self-similarity measured by ΣG. Then corresponding decoy datasets are selected in a way so that $\Sigma F = \Sigma G \Longrightarrow \Sigma S = 0$. But first, you have to determine which is the maximum common diversity that you can achieve with your datasets of actives given the size of the subset of actives you want to select. (In the original MUV datasets, $k = 30$ actives proved to be a reasonable choice.) In order to do so, the maximally diverse subset of with k members has to be selected from each dataset of actives. This can be done using the well established Kennard-Stone algorithm. At the MATLAB prompt type:

```
>>[Rksnn, Sksnn] = spst_ksnn(act, dec, classes, 30, 15000, options)
Output:
Generating datasets of actives with maximum spread. (Kennard-Stone Design)
classe1
classe2
Generating datasets of decoys with minimum separation. (k-Nearest Neighbor Design)
classe1
classe2
Rksnn =
    mode: [1x1 struct]
     act: [1x1 struct]
     dec: [1x1 struct]
Sksnn =
    sigmaG: [2x1 double]
    sigmaF: [2x1 double]
```

This selects a subset of $k = 30$ maximally diverse actives and a subset of $d = 15000$ decoys minimally separated from the actives for each activity class. The struct variable Rksnn provides logical indices into act and dec to identify the selected compounds. The struct variable Sksnn provides ΣG and ΣF values for these selected subsets.

We will use a target value of $\Sigma G = 312$ in this tutorial, since this is the target value of the original MUV datasets. Of course, you could use any value larger than the largest one in Sksnn. In order to set the target value of ΣG for the following final generation of MUV datasets type

```
>>options.targetG=313;    (This is also the Default)
```

D.3.4 Generating MUV Datasets

Now we can generate the actual MUV datasets. The workflow will first adjust all datasets of actives to the common value of $\Sigma G = 312$ using a row exchange algorithm. Then it will employ a genetic algorithm to select a corresponding subset of decoys with $\Sigma F = 312$. The Kennard-Stone design from the previous step is used as a starting design. Type:

```
>>[Rmuv, Smuv] = spst_muv(act, dec, classes, Rksnn, 30, 15000, options)
Output:
Generating datasets of actives with common spread. (MUV Design, Row-Exchange Algorithm)
classe1
Optimized after:2 Iterations. delta(SigmaG)=6.8212e-13 SigmaG=312
classe2
Optimized after:2 Iterations. delta(SigmaG)=0.1 SigmaG=312.1
Generating datasets of decoys with common separation. (MUV Design, Genetic Algorithm)
classe1
Calculating NN-Distances...
Calculating first fitness vector...
```

```
Fitness of starting design:5.67247      20.5051      20.7805      20.8441      20.9285
Start evolution...
1 Iterations. Fitness:0.74127     0.91213      1.0014      1.0025      1.0198, NoChangeCount:0
spst_Fga converged after:1 Iterations. With a final fitness of:0.74127
classe2
Calculating NN-Distances...
Calculating first fitness vector...
Fitness of starting design:0.961133      26.5592      26.6961      26.7561      26.9576
Start evolution...
spst_Fga converged after:0 Iterations. With a final fitness of:0.96113
Rmuv =
    mode: [1x1 struct]
     act: [1x1 struct]
     dec: [1x1 struct]
Smuv =
    sigmaG: [2x1 double]
    sigmaF: [2x1 double]
    sigmaS: [2x1 double]
```

Done! You just generated MUV datasets for the two example datasets. Again, Rmuv constitutes a struct providing logical indices into act and Smuv provides the spatial statistics data of the generated datasets. Let's have a look:

```
>> Smuv.sigmaG
ans =
  312.0000
  312.1000
>> Smuv.sigmaF
ans =
  313.1793
  312.7685
>> Smuv.sigmaS
ans =
    1.1793
    0.6685
```

Both ΣG and ΣF converged to a value very near 312. Accordingly, $\Sigma S \approx 0$ for both datasets.

D.3.5 Exporting the Datasets to SD-Files

In order to use your newly generated MUV datasets for VS validations, you can export them to SD-files on your disk. This is facilitated by the function spst_muv_extractSD.

As arguments it takes the act and dec structs, the just generated Rmuv, the path to the directory of the original SD-Files ('example_dir', in our case), the path where the new MUV SD-Files shall be created, the suffix cell array and the options struct. Before running the script, you have to create the output directory and ensure that you have write access. Let's assume that you chose

```
/home/your_username/muv_sdf
```

as the output directory. Then you would call spst_muv_extractSD as:

```
>> spst_muv_extractSD(act, dec, Rmuv, classes, example_dir, '/home/your_username/muv_sdf', suffix, options)
```

Have fun validating with your brand-new MUV datasets!

Bibliography

(1) Sneader, W. *Drug discovery: a history*; John Wiley & Sons Ltd.: Chichester, UK, 2005.

(2) Boehm, H.-J.; Klebe, G.; Kubinyi, H. *Wirkstoffdesign*; Spektrum Akademischer Verlag GmbH: Heidelberg, Germany, 1996.

(3) Ehrlich, P.; Shiga, K. Farbentherapeutische Versuche zur Trypanosomenerkrankung. *Klin. Wochenschr.* **1904**, *41*, 329–332,362–365.

(4) Steinhilber, D.; Schubert-Zsilavecz, M.; Roth, H. J. *Medizinische Chemie : Targets und Arzneistoffe*; Deutscher Apotheker-Verlag: Stuttgart, Germany, 2005.

(5) *High throughput screening: methods and protocols*; Janzen, W. P., Ed.; Human Press Inc.: Totowa, NJ, 2002.

(6) Boehm, H.; Schneider, G. *Virtual screening for bioactive molecules*; Wiley-VCH: Weinheim, 2000.

(7) Malo, N.; Hanley, J. A.; Cerquozzi, S.; Pelletier, J.; Nadon, R. Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* **2006**, *24*, 167–175.

(8) Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 11473–11478.

(9) Walters, W. P.; Namchuk, M. Designing screens: how to make your hits a hit. *Nat. Rev. Drug Discov.* **2003**, *2*, 259–266.

(10) Mannhold, R.; Kubinyi, H.; Folkers, G. *High throughput screening in drug discovery*; WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2006.

(11) Shoichet, B. K. Screening in a spirit haunted world. *Drug. Discov. Today* **2006**, *11*, 607–615.

(12) National Institutes of Health (NIH), *The Molecular Libraries Initiative, http://mli.nih.gov/mli/ (accessed 14 Feb 2008)*.

(13) Zerhouni, E. Medicine. The NIH roadmap. *Science* **2003**, *302*, 63–72.

(14) National Institutes of Health (NIH), *NIH Roadmap for Medical Research, http://nihroadmap.nih.gov/ (accessed 14 Feb 2008)*.

(15) National Center for Biotechnology Information (NCBI), *PubChem, http://pubchem.ncbi.nlm.nih.gov/ (accessed 14 Feb 2008)*.

(16) Wheeler, D. L. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **2008**, *36*, D13–D21.

(17) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **2002**, *1*, 882–894.

(18) Alvarez, J. C. High-throughput docking as a source of novel drug leads. *Curr. Opin. Chem. Biol.* **2004**, *8*, 365–370.

(19) Stahura, F. L.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* **2005**, *11*, 1189–1202.

(20) Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.* **1894**, *27*, 2985–2993.

(21) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949.

(22) *Fred, 2.2.3*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2006.

(23) Flower, R. J. The development of COX2 inhibitors. *Nat. Rev. Drug Discov.* **2003**, *2*, 179–191.

(24) *The PyMOL molecular graphics system, 1.0r1*; Delano Scientific, LLC.: Palo Alto, CA, 2007.

(25) *Vida3, 3.0.0*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2008.

(26) Klabunde, T.; Hessler, G. Drug design strategies for targeting G-protein-coupled receptors. *ChemBioChem* **2002**, *3*, 928–944.

(27) Birch, P. J.; Dekker, L. V.; James, I. F.; Southan, A.; Cronk, D. Strategies to identify ion channel modulators: current and novel approaches to target neuropathic pain. *Drug. Discov. Today* **2004**, *9*, 410–418.

(28) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.

(29) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(30) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity?. *J. Med. Chem.* **2002**, *45*, 4350–4358.

(31) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.

(32) Ehrlich, P. Über den jetzigen Stand der Chemotherapie. *Ber. Dtsch. Chem. Ges.* **1909**, *42*, 17–47.

(33) Gund, P. Three-dimensional pharmacophoric pattern searching. *Prog. Mol. Subcell. Biol.* **1977**, *5*, 117–143.

(34) *Molecular Operating Environment (MOE), 2007.09*; Chemical Computing Group: Montreal, Canada, 2007.

(35) Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; Wiley-VCH: New York, NY, 2000.

(36) Mason, J.; Morize, I.; Menard, P.; Cheney, D.; Hulme, C.; Labaudiniere, R. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.

(37) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem. Int. Ed. Engl.* **1999**, *38*, 2894–2896.

(38) Baumann, K. An alignment-independent versatile structure descriptor for QSAR and QSPR based on the distribution of molecular features. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 26–35.

(39) Stiefl, N.; Baumann, K. Mapping property distributions of molecular surfaces: algorithm and evaluation of a novel 3D quantitative structure-activity relationship technique. *J. Med. Chem.* **2003**, *46*, 1390–1407.

(40) Jaquard, P. Distribution de la flore alpine dans le bassin des dranses et dans quelques regions voisines. *Bull. Soc. Vaud. Sci. Nat.* **1901**, *37*, 241–272.

(41) Tanimoto, T. T. *IBM internal report 17th Nov.*; Technical Report, 1957.

(42) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233–245.

(43) Godden, J. W.; Furr, J. R.; Xue, L.; Stahura, F. L.; Bajorath, J. Molecular similarity analysis and virtual screening by mapping of consensus positions in binary-transformed chemical descriptor spaces with variable dimensionality. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 21–29.

(44) Roche, O. et al. Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *J. Med. Chem.* **2002**, *45*, 137–142.

(45) Geppert, H.; Horvath, T.; Gärtner, T.; Wrobel, S.; Bajorath, J. Support-vector-machine-based ranking significantly improves the effectiveness of similarity searching using 2D fingerprints and multiple reference compounds. *J. Chem. Inf. Model.* **2008**, *48*, 742–746.

(46) Hsieh, J.-H.; Wang, X.; Teotico, D.; Golbraikh, A.; Tropsha, A. Differentiation of AmpC beta-lactamase binders vs. decoys using classification kNN QSAR modeling and application of the QSAR classifier to virtual screening. *J. Comput.-Aided Mol. Des.* **2008**.

(47) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.

(48) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.

(49) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.

(50) Brown, R.; Martin, Y. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.

(51) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.

(52) Reid, D.; Sadjad, B.; Zsoldos, Z.; Simon, A. LASSO-ligand activity by surface similarity order: a new tool for ligand based virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *6-7*, 479–487.

(53) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74–82.

(54) Stiefl, N.; Watson, I.; Baumann, K.; Zaliani, A. ErG: 2D Pharmacophore descriptions for scaffold hopping. *J. Chem. Inf. Model.* **2006**, *46*, 208–220.

(55) Schulz-Gasch, T.; Stahl, M. Scoring functions for protein-ligand interactions: a critical perspective. *Drug Discov. Today: Technol.* **2004**, *1*, 231–239.

(56) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of automated docking programs as virtual screening tools. *J. Med. Chem.* **2005**, *48*, 962–976.

(57) Pearlman, D.; Charifson, P. Improved scoring of ligand-protein interactions using OWFEG free energy grids. *J. Med. Chem.* **2001**, *44*, 502–511.

(58) Bender, A.; Glen, R. C. A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369–1375.

(59) Sheridan, R.; Singh, S.; Fluder, E.; Kearsley, S. Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J. Chem. Inf. Model.* **2001**, *41*, 1395–1406.

(60) Truchon, J.-F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

(61) Klon, A.; Glick, M.; Davies, J. Application of machine learning to improve the results of high-throughput docking against the HIV-1 protease. *J. Chem. Inf. Model.* **2004**, *44*, 2216–2224.

(62) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.

(63) Cleves, A.; Jain, A. Robust ligand-based modeling of the biological targets of known drugs. *J. Med. Chem.* **2006**, *49*, 2921–2938.

(64) Pham, T.; Jain, A. Parameter estimation for scoring protein-ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49*, 5856–5868.

(65) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759–4767.

(66) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.

(67) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(68) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(69) *MDL Drug Data Report (MDDR)*; Symyx Technologies, Inc.: Santa Clara, CA, 2005.

(70) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information. *J. Med. Chem.* **2005**, *48*, 7049–7054.

(71) Irwin, J. J.; Shoichet, B. K. ZINC-a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

(72) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793–806.

(73) Good, A. C.; Hermsmeier, M. A.; Hindle, S. Measuring CAMD technique performance: a virtual screening case study in the design of validation experiments. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 529–536.

(74) Good, A. C.; Hermsmeier, M. A. Measuring CAMD technique performance. 2. How "druglike" are drugs? implications of random test set selection exemplified using druglikeness classification models. *J. Chem. Inf. Model.* **2007**, *47*, 110–114.

(75) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection?. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169–178.

(76) Vogt, M.; Godden, J. W.; Bajorath, J. Bayesian interpretation of a distance function for navigating high-dimensional descriptor spaces. *J. Chem. Inf. Model.* **2007**, *47*, 39–46.

(77) Vogt, M.; Bajorath, J. Introduction of an information-theoretic method to predict recovery rates of active compounds for bayesian in silico screening: theory and screening trials. *J. Chem. Inf. Model.* **2007**, *47*, 337–341.

(78) Weiniger, D. In *Encyclopedia of computational chemistry, Vol. 1*; Schleyer, P. V. R., Allinger, N., Clark, T., Gasteiger, J., Kollman, P., Schaefer, H., Eds.; Wiley & Sons, Ltd.: New York, NY, 1998; Chapter Combinatorics of small molecular structures, pp 425–430.

(79) Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855–861.

(80) *CTFile formats*; Symyx Technologies, Inc.: Santa Clara, CA, 2005.

(81) *3D structure generator CORINA: generation of high-quality three-dimensional molecular models*; Molecular Networks GmbH Computerchemie: Erlangen, Germany, 2006.

(82) Jolliffe, I. *Principal component analysis*; Springer-Verlag, New York, N.Y., 2002.

(83) *BABEL3, 2.2*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2006.

(84) *Filter, 2.2.1*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2007.

(85) De Aguiar, P.; Bourguignon, B.; Khots, M.; Massart, D.; Phan-Than-Luu, R. D-optimal designs. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 199–210.

(86) Johnson, M. E.; Nachtsheim, C. J. Some guidelines for constructing exact D-optimal designs on convex design spaces. *Technometrics* **1983**, *25*, 271–277.

(87) Olsson, I.-M.; Gottfries, J.; Wold, S. D-optimal onion designs in statistical molecular design. *Chemom. Intell. Lab. Syst.* **2004**, *73*, 37–46.

(88) Olsson, I.-M.; Gottfries, J.; Wold, S. Controlling coverage of d-optimal onion designs and selections. *J. Chemom.* **2004**, *18*, 548–557.

(89) Scott, D. W.; Thompson, J. R. Probability density estimation in higher dimensions. *Interface: Computer science and statistics, Proceedings of the 15th Symposium*, North-Holland: Amsterdam, The Netherlands, 1983; pp 173–179.

(90) *MATLAB 7*; The Mathworks: Natick, MA, 2006.

(91) Atkinson, A. C.; Donev, A. N. *Optimum experimental designs*; Oxford University Press: Oxford, UK, 1992.

(92) Box, G. E.; Hunter, S.; Hunter, W. *Statistics for experimenters: desing, innovation, and discovery*, 2nd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, 2005.

(93) Upton, G. J. G.; Fingleton, B. *Spatial dana analysis by example*; Wiley & Sons Ltd.: New York, N.Y., 1985.

(94) Fortin, M.-J.; Dale, M. R. T. *Spatial analysis: a guide for ecologists*; Cambridge University Press: Cambridge, UK, 2005.

(95) Breimann, L. *Using convex pseudo-data to improve prediction accuracy.*; Technical Report, 1998.

(96) Kohonen, T. *Self-organizing maps*; Springer-Verlag New York, Inc.: Secaucus, NJ, USA, 1997.

(97) Vesanto, J. M.Sc. thesis, Helsinki University of Technology, 1997.

(98) Vesanto, J. SOM-based data visualization methods. *Intelligent Data Analysis* **1999**, *2*, 111–126.

(99) Vesanto, J.; Alhoniemi, E. Clustering of the self-organizing map. *IEEE T. Neural. Networ.* **2000**, *11*, 586–600.

(100) Kaski, S.; Lagus, K. Comparing self-organizing maps. *Proceedings of ICANN96, International Conference on Artificial Neural Networks, Bochum, Germany, July 16-19*, Berlin, 1996; pp 809–814.

(101) Kaski, S. Data exploration using self-organizing maps. *Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82* **1997**.

(102) Vesanto, J.; Himberg, J.; Alhoniemi, E.; Parhankangas, J. Self-organizing map in matlab: the SOM toolbox. *Proceedings of the Matlab DSP Conference 1999, Espoo, Finland*, 1999; p 3540.

(103) Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **1904**, *15*, 72–101.

(104) Edgington, E. S. *Randomization tests*; Marcel Dekker, Inc., New York, NY, 1980.

(105) National Center for Biotechnology Information (NCBI), *PubChem Power User Gateway (PUG), http://pubchem.ncbi.nlm.nih.gov/pug/pughelp.html (accessed 9 Jul 2008)*.

(106) Zhou, Y.; Zhou, B.; Chen, K.; Yan, S. F.; King, F. J.; Jiang, S.; Winzeler, E. A. Large-scale annotation of small-molecule libraries using public databases. *J. Chem. Inf. Model.* **2007**, *47*, 1386–1394.

(107) Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K. High-throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.* **2005**, *1*, 146–148.

(108) Crisman, T. J.; Parker, C. N.; Jenkins, J. L.; Scheiber, J.; Thoma, M.; Kang, Z. B.; Kim, R.; Bender, A.; Nettles, J. H.; Davies, J. W.; Glick, M. Understanding false positives in reporter gene assays: in silico chemogenomics approaches to prioritize cell-based HTS data. *J. Chem. Inf. Model.* **2007**, *47*, 1319–1327.

(109) Auld, D. S.; Southall, N. T.; Jadhav, A.; Johnson, R. L.; Diller, D. J.; Simeonov, A.; Austin, C. P.; Inglese, J. Characterization of chemical libraries for luciferase inhibitory activity. *J. Med. Chem.* **2008**, *51*, 2372–2386.

(110) Simeonov, A.; Jadhav, A.; Thomas, C. J.; Wang, Y.; Huang, R.; Southall, N. T.; Shinn, P.; Smith, J.; Austin, C. P.; Auld, D. S.; Inglese, J. Fluorescence spectroscopic profiling of compound libraries. *J. Med. Chem.* **2008**, *51*, 2363–2371.

(111) Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P. A high-throughput screen for aggregation-based inhibition in a large compound library. *J. Med. Chem.* **2007**, *50*, 2385–2390.

(112) Motulsky, H. *Analyzing data with graphpad prism.*; Graphpad Software Inc., 1999.

(113) *GraphPad Prism, 4*; GraphPad Software, Inc.: San Diego, CA, 2003.

(114) Pearce, B.; Sofia, M.; Good, A.; Drexler, D.; Stock, D. An empirical process for the design of high-throughput screening deck filters. *J. Chem. Inf. Model.* **2006**, *46*, 1060–1068.

(115) Merkwirth, C.; Mauser, H.; Schulz-Gasch, T.; Roche, O.; Stahl, M.; Lengauer, T. Ensemble methods for classification in cheminformatics. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1971–1978.

(116) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *22*, 4673–4680.

(117) Chenna, R.; Sugawara, H.; Koike, T.; Lopez, R.; Gibson, T. J.; Higgins, D. G.; Thompson, J. D. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **2003**, *31*, 3497–3500.

(118) Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947–2948.

(119) *SciFinder Scholar, 2007*; Chemical Abstracts Service: Columbus, OH, 2007.

(120) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906.

(121) *Prous Drugs of the Future*; Prous Science: Philadelphia, PA, 2008.

(122) Sigma-Aldrich, *Chemistry Product Catalog, http://www.sigmaaldrich.com (accessed Mar 7, 2008)*.

(123) *MDL MACCS keys*; Symyx Technologies, Inc.: Santa Clara, CA, 2005.

(124) Mandel, J. Use of the singular value decomposition in regression analysis. *Amer. Statistician* **1982**, *36*, 15–24.

(125) Kennard, R. W.; Stone, L. A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137–148.

(126) Clark, R. OptiSim: an extended dissimilarity selection method for finding diverse representative subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181–1188.

(127) Renner, S.; Schneider, G. Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem* **2006**, *1*, 181–185.

(128) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(129) Xu, Y.-J.; Johnson, M. Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *J. Chem. Inf. Model.* **2002**, *42*, 912–926.

(130) Xu, J. A new approach to finding natural chemical structure classes. *J. Med. Chem.* **2002**, *45*, 5311–5320.

(131) McGregor, M.; Pallai, P. Clustering of large databases of compounds: using MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.

(132) Rohrer, S. G.; Baumann, K. Impact of benchmark data set topology on the validation of virtual screening methods: exploration and quantification by spatial statistics. *J. Chem. Inf. Model.* **2008**, *48*, 704–718.

(133) Vogt, M.; Bajorath, J. Introduction of a generally applicable method to estimate retrieval of active molecules for similarity searching using fingerprints. *ChemMedChem* **2007**, *2*, 1311–1320.

(134) Schmuker, M.; Schneider, G. Processing and classification of chemical data inspired by insect olfaction. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 20285–20289.

(135) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and management of chemical properties in CACTVS: an extensible networked approach toward modularity and compatibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109–116.

(136) *QikProp, 3.0*; Schrödinger, LLC: New York, NY, 2007.

(137) Irwin, J. Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 193–199.

(138) Jacoby, E. et al. Key aspects of the novartis compound collection enhancement project for the compilation of a comprehensive chemogenomics drug discovery screening collection. *Curr. Top. Med. Chem.* **2005**, *5*, 397–411.

(139) Jain, A.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133–139.

(140) *PubChem Standardization Service http://pubchem.ncbi.nlm.nih.gov/standardize-/standardize.cgi (accessed 11 Aug 2008)*; National Center for Biotechnology Information (NCBI).

(141) Koong, A. C.; Feldman, D. E. *Methods to identify inhibitors of the unfolded protein response.*, PCT Int. Appl., WO 2007101225, 2007.

(142) Koong, A. C.; Feldman, D. E. *Inhibitors of the unfolded protein response and endonuclease IRE1 and methods for therapeutic use.*, PCT Int. Appl., WO 2007101224, 2007.

(143) Savel'ev, V. L.; Pryanishnikova, N. T.; Zagorevskii, V. A.; Chernyakova, I. V.; Artamonova, O. S.; Shavyrina, V. V.; Malysheva, L. I. Synthesis and pharmacological activity of 4H-(1)-benzopyrano-(3,4-d)-imidazol-4-ones. *Farm. Zh.* **1983**, *17*, 697–700.

(144) Tsizin, Y. S. Heterocyclic quinones. XXIII. 2-chloroquinolinequinones. *Khimiya Geterotsiklicheskikh Soedinenii* **1973**, 1700–1704.

(145) Malon, M.; Travnicek, Z.; Marysko, M.; Zboril, R.; Maslan, M.; Marek, J.; Dolezal, K.; Rolcik, J.; Krystof, V.; Strnad, M. Metal complexes as anticancer agents 2. Iron(III) and copper(II) bio-active complexes with N6-benzylaminopurine derivatives. *Inorg. Chim. Acta* **2001**, *323*, 119–129.

(146) Klokol, G. V.; Krivokolysko, S. G.; Dyachenko, V. D.; Litvinov, V. P. Aliphatic aldehydes in the synthesis of condensed 4-alkyl(cycloalkyl)-2-amino-3-cyano-4h-pyrans. *Chemistry of Heterocyclic Compounds (New York), (Translation of Khimiya Geterotsiklicheskikh Soedinenii)* **1999 (2000)**, *35*, 1183–1186.

(147) Dyachenko, V. D.; Krivokolysko, S. G.; Litvinov, V. P. Synthesis and alkylation of 3-cyano-4,7,7-trimethyl-2-thioxo-1,2,3,4,5,6,7,8-octahydroquinolin-5-one. *Russ. J. Org. Chem.* **1998**, *34*, 707–711.

(148) Strakovs, A.; Tonkikh, N. N.; Petrova, M.; Ryzhanova, K. V.; Palitis, E. The reaction of 2-aminoethyl-and 3-aminopropyl-substituted heterocycles with 2-formyl-1,3-cyclanediones and 4-oxo-3,1-benzoxazines. *Chemistry of Heterocyclic Compounds (New York, NY)(Translation of Khimiya Geterotsiklicheskikh Soedinenii)* **2002**, *38*, 449–455.

(149) Fu, J.; Hou, D.; Kamboj, R.; Kodumuru, V.; Pokrovskaia, N.; Raina, V.; Sun, S.; Sviridov, S.; Zhang, Z. *Preparation of aminothiazole derivatives as human stearoyl-CoA desaturase inhibitors.*, PCT Int. Appl., WO 207130075, 2007.

(150) Alexandrova, L. A.; Jasko, M. V.; Belobritskaya, E. E.; Chudinov, A. V.; Mityaeva, O. N.; Nasedkina, T. V.; Zasedatelev, A. S.; Kukhanova, M. K. New triphosphate conjugates bearing reporter groups: labeling of DNA fragments for microarray analysis. *Bioconjug. Chem.* **2007**, *18*, 886–893.

(151) Tang, P. C.; Sun, L.; Shawver, L. K.; Hirth, K. P.; Fong, A. *Methods for treating diseases and disorders related to unregulated angiogenesis and/or vasculogenesis.*, U.S. Patent, US 6,147,106, 2005.

(152) Howlett, A. R.; Rice, A.; Moshinsky, D.; Hammarsten, O. *A preparation of sulfonamide substituted indolinones, useful as inhibitors of dna dependent protein kinase (DNA-PK).*, U.S. Pat. Appl. Publ., US 2004266843, 2004.

(153) Tang, P. C.; Liang, C.; Miller, T.; Lipson, K. E. *Preparation of 5-sulfonamido-substituted indolinone compounds as protein kinase inhibitors.*, U.S. Pat. Appl. Publ., US 2004204407, 2004.

(154) Tang, P. C.; Sun, L.; Miller, T. A.; Liang, C.; Tran, N. M.; Nguyen, A. T.; Nematalla, A. *3-methylidenyl-2-indolinone modulators of protein kinase.*, PCT Int. Appl., WO 2000008202, 2000.

(155) Nogales, D. F.; Lightner, D. A. Synthesis of a [13CO2H]-labeled bilirubin. *J. Labelled Comp. Radiopharm.* **1994**, *34*, 453–62.

(156) Meltola, N. J.; Soini, A. E.; Haenninen, P. E. Syntheses of novel Dipyrrylmethene-BF2 dyes and their performance as labels in two-photon excited fluoroimmunoassay. *J. Fluoresc.* **2004**, *14*, 129–138.

(157) Meltola, N.; Soini, A. *Two-photon absorbing dipyrromethene boron difluoride dyes and their applications.*, PCT Int. Appl., WO 2003005030, 2003.

(158) Black, D. S. Product class 13: 1H-pyrroles. *Science of Synthesis* **2002**, *9*, 441–552.

(159) Bobal, P.; Lightner, D. A. An inexpensive, selective procedure for oxidizing methyl to formyl pyrroles. *J. Heterocycl. Chem.* **2001**, *38*, 1219–1221.

(160) Kadota, H.; Sehata, M.; Maruyama, K.; Sakai, K. *Preparation of pyrrole-2-carboxylic acids from aminomalonates and diketones.*, Jpn. Kokai Tokkyo Koho, JP 2001288167, 2001.

(161) Tu, B.; Wang, C.; Ma, J. Improved synthesis of symmetrical dipyrromethenes. *Org. Prep. Proced. Int.* **1999**, *31*, 349–352.

(162) Cheng, L. J.; Lightner, D. A. Synthesis of cyanopyrroles. *Synthesis* **1999**, 46–48.

(163) Bereziin, M. V.; Semeikin, A. S.; V'ugin, A. I.; Krestov, G. A. Thermochemistry of substituted pyrroles. *Izv. Akad. Nauk SSSR, Ser. Khim.* **1993**, 495–499.

(164) Ma, J.; Chen, Q.; Cheng, L.; Wang, C.; Jin, S. Intramolecular hydrogen bonding in symmetric dipyrrylmethanes. *Spectrosc. Lett.* **1995**, *28*, 223–230.

(165) Ganske, J. A.; Pandey, R. K.; Postich, M. J.; Snow, K. M.; Smith, K. M. Some mercuration reactions of substituted pyrroles. *J. Org. Chem.* **1989**, *54*, 4801–4807.

(166) Paine, I., John B.; Dolphin, D. Pyrrole chemistry. An improved synthesis of ethyl pyrrole-2-carboxylate esters from diethyl aminomalonate. *J. Org. Chem.* **1985**, *50*, 5598–5604.

(167) Paine, I., John B.; Chang, C. K.; Dolphin, D. The synthesis of porphyrins via dipyrromethenes. *Heterocycles* **1977**, *7*, 831–838.

(168) Mironov, A. F.; Miroshnichenko, L. D.; Evstigneeva, R. P.; Preobrazhenskii, N. A. Infrared spectra of pyrroles and dipyrrylmethanes. *Khimiya Geterotsiklicheskikh Soedinenii* **1965**, 74–80.

(169) Hayes, A.; Kenner, G. W.; Williams, N. R. Pyrroles and related compounds. I. Synthesis of some unsymmetrical pyrrolylmethylpyrroles (pyrromethanes). *J. Chem. Soc.* **1958**, 3779–3788.

(170) Udelhofen, J. H. Ph.D. thesis, Iowa State Coll., Ames, IA, 1959.

(171) Wright, G. C. Ph.D. thesis, Univ. of Delaware, Newark, DE, 1959.

(172) Fischer, H.; Sus, O.; Weilguny, F. G. Curtius degradation in the pyrrole series. I. *Justus Liebigs Annalen der Chemie* **1930**, *481*, 159–192.

(173) Kirino, O.; Yamamoto, S.; Kato, T. Structure-activity study of fungicidal N-benzoylanthranilates. part II. Structure-activity relationships of fungicidal N-benzoylanthranilic esters. *Agric. Biol. Chem.* **1980**, *44*, 2149–2153.

(174) Sangapure, S. S.; Agasimundin, Y. S. Studies in benzofurans: part III. Synthesis and reactions of 2-alkyl- or 2-aryl-3,4-dihydro-4-oxobenzofuro[3,2-d]pyrimidines and 4-thio analogs. *Indian J. Chem., Sect. B: Org. Chem. Incl. Med. Chem.* **1978**, *16B*, 627–629.

(175) Zhao, X.; Lan, Y.; Wu, C. *Method for synthesizing N-methylpiperazine-substituted aniline.*, Faming Zhuanli Shenqing Gongkai Shuomingshu, CN 101168532, 2008.

(176) Busch, B. B.; Flatt, B. T.; Gu, X. H.; Martin, R.; Mohan, R.; Nyman, M. C.; Schweiger, E.; Stevens, J., William C.; Wang, T. L.; Xie, Y. *Pyrazole based LXR modulators and their preparation, pharmaceutical compositions and use in the treatment of diseases.*, PCT Int. Appl., WO 2007002559, 2007.

(177) Angell, R. M.; Aston, N. M.; Bamborough, P.; Bamford, M. J.; Cockerill, G. S.; Flack, S. S.; Laine, D. I.; Walker, A. L. *Preparation of biphenylcarboxamides as p38 kinase inhibitors.*, PCT Int. Appl., WO 2003033483, 2003.

(178) Angell, R. M.; Aston, N. M.; Bamborough, P.; Bamford, M. J.; Cockerill, G. S.; Merrick, S. J.; Walker, A. L. *Preparation of 5'-carbamoyl-1,1'-biphenyl-4-carboxamides as p38 kinase inhibitors.*, PCT Int. Appl., WO 2003032980, 2003.

(179) Howard, H. R. *Preparation of heterocyclic carboxamides as 5-HT1 agonists or antagonists.*, Eur. Pat. Appl., EP 957099, 1999.

(180) Halazy, S.; Lamothe, M.; Jorand-Lebrun, C. *Preparation of piperazines as antagonists of 5-HT1A, 5-HT1D, and 5-HT1B receptors.*, PCT Int. Appl., WO 9842692, 1998.

(181) Setoi, H.; Ohkawa, T.; Zenkoh, T.; Sawada, H.; Sawada, Y.; Oku, T. *Preparation of benzamide derivatives having a vasopressin antagonistic activity.*, PCT Int. Appl., WO 9824771, 1998.

(182) Howard, H. R.; Segelstein, B. E. *Preparation of 3-(piperazinophenyl)acrylamides and analogs as 5-HT1 receptor ligands.*, Eur. Pat. Appl., EP 810220, 1997.

(183) Halazy, S.; Lamothe, M. *Arylpiperazine cyclic amine derivatives as 5HT1D receptor antagonists.*, PCT Int. Appl., WO 9714689, 1997.

(184) Lamothe, M.; Perez, M.; Colovary-Gotteland, V.; Halazy, S. A simple one-pot preparation of N,N'-unsymmetrical ureas from N-Boc protected primary anilines and amines. *Synlett* **1996**, 507–508.

(185) Nishimura, T.; Oyama, H.; Yamamura, H.; Morita, T.; Matsumoto, K.; Watanabe, T. *Pyrazolylpyrimidine fungicides.*, Jpn. Kokai Tokkyo Koho, JP 54147921, 1979.

(186) Agarwal, A.; Louise-May, S.; Thanassi, J. A.; Podos, S. D.; Cheng, J.; Thoma, C.; Liu, C.; Wiles, J. A.; Nelson, D. M.; Phadke, A. S.; Bradbury, B. J.; Desh-

pande, M. S.; Pucci, M. J. Small molecule inhibitors of E. coli primase, a novel bacterial target. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 2807–2810.

(187) Avdeenko, A. P.; Konovalova, S. A.; Il'chenko, A. Y.; Glinyanaya, N. M. Halogenation of N-substituted p-quinone imines and p-quinone oxime esters: III. Regioselectivity in the halogenation of N-aroyl(arylsulfonyl)oxyimino-2,5-cyclohexadienones. *Russ. J. Org. Chem.* **2006**, *42*, 56–65.

(188) Avdeenko, A. P.; Zhukova, S. A.; Glinyanaya, N. M.; Konovalova, S. A. Halogenation of 4-aroyl(arenesulfonyl)oximino-2,6(3,5)-dimethylcyclohexa-2,5-dien-1-ones. *Russ. J. Org. Chem.* **1999**, *35*, 560–571.

(189) Samee, W.; Ungwitayatorn, J.; Matayatsuk, C.; Pimthon, J. 3D-QSAR studies on phthalimide derivatives as HIV-1 reverse transcriptase inhibitors. *Sci. Asia* **2004**, *30*, 81–88.

(190) Ungwitayatorn, J.; Matayatsuk, C.; Sothipatcharasai, P. Quantitative structure-activity relationship study on phthalimide derivatives as HIV-1 reverse transcriptase inhibitors. *Sci. Asia* **2001**, *27*, 245–250.

(191) Lafon, L. *N-propargyl piperonylamine.*, Ger. Offen., DE 2712432, 1977.

(192) Gompper, R.; Sramek, M. Anthraquinone and napthacenequinone from napthoquinones or 1,4-dihydroxyanthraquinone and 4-dimethylamino-1,1,2-trimethoxybutadiene. *Synthesis* **1981**, 649–50.

Symbols and Abbreviations

$std_{top}(FoM)$	Standard deviation component caused by dataset topology
ACE	Angiotensin Converting Enzyme
AChE	Acetylcholineesterase
AID	Assay Identity
ANN	Artificial Neural Network
CID	Compound Identity
Conf. Itv.	Confidence Interval
COX	Cyclooxygenase
COX2	Cyclooxygenase 2
D1 Rec.	Dopamine D1 Receptor
DUD	Directory of Useful Decoys
EF	Enrichment Factor
Eph rec. A4	Eph Receptor A4
ER-α-Coact.	Estrogen Receptor-α Coactivator
ER-β-Coact.	Estrogen Receptor-β Coactivator
FAK	Focal Adhesion Kinase

FoH	Frequency of Hits
FXIa	Factor XIa
FXIIa	Factor XIIa
GPCR	G-Protein Coupled Receptor
HIV	Human Immunodeficiency Virus
HSP90	Heat Shock Protein 90
HTS	High-Throughput Screening
LBVS	Ligand Based Virtual Screening
MDDR	MDL Drug Data Report
MLI	Molecular Libraries Initiative
MLSCN	Molecular Libraries Screening Centers Network
MOE	Molecular Operating Environment
MUV	Maximum Unbiased Validation
NIH	National Institutes of Health
PA	Potential Actives
PCA	Principal Components Analysis
PD	Potential Decoys
PDB	Protein Data Bank
PKA	Protein Kinase A
PKC	Protein Kinase C
PPI	Protein Protein Interaction

PUG	PubChem Power User Gateway
qHTS	Quantitative High-Throughput Screening
Rec. Tyr. Kinase	Receptor Tyrosine Kinase
ROC	Receiver Operating Characteristic
RT	Reverse Transcriptase
RTR	Retrieval Rate
S1P1	Sphingosine-1-Phosphate Receptor 1
S1P2	Sphingosine-1-Phosphate Receptor 2
SBVS	Structure Based Virtual Screening
SF1	Nuclear Receptor Steroidogenic Factor 1
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TAC	Tested Assay Count
UID	Unique Identity
VS	Virtual Screening
wAAC	Weighted Active Assay Count
ZINC	ZINC Is Not Commercial
d_E	Euclidean Distance
EC_{50}	Half maximal effective concentration
λ	Goodness of a VS ranking
N	Number of decoys in a validation set

R_a	Fraction of actives in a validtation set
AUC_{ROC}	Area under the Receiver Operating Characterisitc curve
$RTR_{1\%}$	Retrieval Rate at 1 percent of the ranking
$\%SI$	Percent Sequence Identity
σ^2	Variance
$\sigma^2_{bt,i}$	Bootstrapping variance component
ΣS	Dataset Clumping
σ^2_{top}	Variance component introduced by dataset topology
T_c	Tanimoto-Jaquard coefficient

Index

AUC_{ROC}, 17, 31

EC_{50}, 67

R_a, 32

Activity space, 21

Analogue bias, 21, 24, 28

Artificial enrichment, 20, 24, 28

Autoscaling, 28, 77

Basic tasks of VS validation, 15

Benchmark Datasets, 19

Benchmarking, 15, 62

CACTVS, 97

Cellular Growth and Proliferation Assays, 4

Chemical similarity, 10

Chemical Space, 21

ClustalW, 72

Clustering, 23

Concentrated, 33

Confirmatory screen, 67

Correlation, 41

Cumulative probability distribution, 34

D-optimal Design, 29

Decoys, 15, 20

Descriptor, 10

Descriptor space, 23

descriptor space, 13

Deterministic Design, 29

Dispersed, 33

Distance measures, 13

Docking, 8

Dose-response relationship, 67

DrugBank, 75

DUD, 20, 97

Dumb descriptor, 20

E-Utilities, 72

Eigenvalues, 28

Embedding, 75

Empty space function, 34

Enrichment factors, 17

Enzymatic Assays, 4

Euclidean distance, 13

Event, 33

F(t), 34

Fingerprints, 11

FoH, 71, 104

FoM, 84

Frequency of hits, 71

Functional Cell-based Assays, 4

G(t), 34
Guide tree, 72

Hert-Willett datasets, 19, 25
Histogram, 72
HTS, 62, 72, 88
HTS Assay Formats, 4

λ, 31
LBVS, 6
Lock and Key, 6

MACCS, 77
MAX-rule data fusion, 14
MDDR, 19, 20, 25, 75
Minimum Diversity Design, 29
MLSCN, 5
MOE, 27
MOE descriptors, 27, 77
Molecular Libraries Initiative, 5, 65
Molecular Libraries Screening Centers Network, 5
MUV, 62, 67

N, 31
Nearest neighbor function, 34

Onion design, 29

Patchiness, 23
Patchy, 33
PCA, 77
PCBioAssay, 5, 65
PCCompound, 5, 65
PDB, 19
Perl, 72
Pharmacophore, 10
Points, 32
Posing, 8
Primary screen, 67
Property descriptors, 11, 27
Prous Drugs of the Future, 75
PubChem, 62, 65
PubChem Power User Gateway, 67
PUG, 67

qHTS, 74
Query, 15

Ranking, 15
Retrieval rate, 17
Roadmap, 5, 65
Row exchange algorithm, 29, 82
RTR, 17, 20
$RTR_{1\%}$, 17, 30

SBVS, 6, 19
Scaffold, 20, 23
Scoring, 8
Self-similarity, 23
SESP, 77
ΣS, 35
Sigma-Aldrich Catalogue, 75
σ^2, 32

$\sigma^2_{bt,i}$, 32

σ^2_{top}, 32

Simple descriptors, 28

Spatial Statistics, 32

Spread, 23

Standard deviation, 32

Structure based virtual screening, 6

Suitability testing, 15

SVD, 77

Tanimoto-Jaquard coefficient, 13

Topological descriptors, 11

Topology, 23, 32

Topology prototypes, 29

Validation, 14

Validation set, 15

Variance, 28

Variance decomposition, 31

Weights, 72

ZINC, 19, 97

Curriculum Vitae

Sebastian Georgios Rohrer

Campus:
Technische Universität Carolo-Wilhelmina zu Braunschweig
Beethovenstr. 55
Telefon: +49-(0)531-3912797

Institut für Pharmazeutische Chemie
38106 Braunschweig
E-mail: s.rohrer@tu-bs.de

Privat:
Beethovenstr. 2
Telefon: +49-(0)931-7809028

38106 Braunschweig
Mobil: +49-(0)176-24079017

Persönliche Angaben

Geburtsdatum:	02. April 1979
Geburtsort:	Nürnberg

Ausbildung

Diplom-Biologe

11/1999 - 11/2004	Bayerische Julius-Maximilians-Universität Würzburg *Hauptfach:* Biotechnologie *Nebenfächer:* Bioinformatik, Physiologische Chemie Abschlussnote: 1,0 (mit Auszeichnung)

Abitur

09/1989 - 06/1998	Gymnasium Donauwörth *Leistungskurse*: Mathematik, Chemie Abschlussnote: 1,5

Zivildienst

09/1998 - 10/1999	Johanniter Unfallhilfe e.V. Regionalverband Nordschwaben-Augsburg Geschäftsstelle Donauwörth

Wissenschaftliche Erfahrung

Forschung

Doktorarbeit

seit 02/2005	Technische Universität Carolo-Wilhelmina zu Braunschweig Institut für Pharmazeutische Chemie *und* Bayerische Julius-Maximilians-Universität Würzburg Institut für Pharmazeutische Chemie (Umzug der Gruppe von Prof. Baumann nach Braunschweig 09/2006) Entwicklung eines mathematischen Formalismus zur räumlichen Analyse chemischer Datensätze und deren Einfluss auf die Validierung von Techniken des Virtuellen Screenings. Analyse und Design großer Datensätze für unverzerrte Validierexperimente. Integration von Inhibitionsprofilen, Targetsequenzen und Enzymkinetikdaten zur Steigerung der Verlässlichkeit von Bioaktivitätsdaten. Computersupport für mehrere medizinisch-chemische Kooperationsprojekte. Eigenständige Koordination der Projekte mit beteiligten Synthesechemikern. Resultate: Entdeckung einer neuen Klasse von Cysteinproteaseinhibitoren virulenter Protozoen (Patent eingereicht), Entdeckung eines Inhibitors des Nipah-Virus Fusionsproteins.

Diplomarbeit

02/2004 - 10/2004	Aventis Pharma Deutschland, Department of Bioinformatics, Frankfurt a. M. Entwicklung und Anwendung einer Software zur Vorhersage von post-translationalen Modifikationsstellen in Peptidhormon-Precursormolekülen Betreuer: Dr. Eva Jung, Prof. Dr. Klaus-Peter Koller (Aventis) Prof. Dr. Thomas Dandekar (Universität Würzburg)

Praktikum

02/2003 - 05/2003	Venetian Institute for Molecular Medicine (VIMM), Padua, Italien ERASMUS-Auslandssemester Planung und Durchführung von Experimenten im Rahmen eines Projektes zur Charakterisierung der Interaktion von VIP (VacA Interacting Protein) und N-WASP (Neural Wiskott Aldrich Syndrome Protein)

Praktikum

08/2002 - 10/2002	Octagene Biomedical Laboratories, Martinsried Etablierung und Optimierung eines Protokolls zur Aufreinigung des humanen Progesteron Rezeptors A (hPR-A)

Lehre

Praktikum und Seminar für Studenten der Pharmazie im 2. Semester: "Quantitative Analyse von Arznei-, Hilfs- und Schadstoffen"

Praktikum für Studenten der Pharmazie im 5. Semester: "Biochemische Methoden"

Wahlpflichtpraktikum für Studenten der Pharmazie im 7. Semester: "Anwendung und Vergleich von Programmen zum Protein-Ligand Docking"

Sonderaufgabe für Studenten der Pharmazie im 8. Semester: "Case-studies in computer aided drug discovery"

Publikationen

Originalpublikationen in Fachjournalen

(1) Rohrer S.G., Baumann K.

The impact of benchmark dataset topology on the validation of virtual screening methods: exploration and quantification by spatial statistics

J. Chem. Inf. Model., **2008**, 48, 704-18

(2) Degel B., Staib P., Rohrer S., Scheiber J., Martina E., Büchold C., Baumann K., Morschhäuser J., Schirmeister T.

Cis-Configured Aziridines Are New Pseudo-Irreversible Dual-Mode Inhibitors of Candida albicans Secreted Aspartic Protease 2

ChemMedChem, **2008**, 3, 302-315

(3) Rohrer S.G., Baumann K.

Maximum Unbiased Validation (MUV) Datasets for Virtual Screening Based on PubChem Bioactivity Data

J. Chem. Inf. Model. (in press)

Vorträge auf Fachkongressen

(1) Rohrer, S.G., Baumann K.

Maximum Unbiased Validation (MUV) Datasets for Benchmarking of Ligand-Based Virtual Screening Techniques

Jahrestagung der Deutschen Pharmazeutischen Gesellschaft, Bonn, 10/2008

(2) Rohrer, S.G.

Exploring Benchmark Dataset Bias in Ligand Based Virtual Screening

Informa Life Sciences: Design and Synthesis of Quality Compound Libraries, München, 12/2007

(3) Rohrer, S.G.

Fair Play in Virtual Screening: Ruling Out Database Composition as a Critical Factor in the Validation of Ligand Based Virtual Screening Methods

20. Darmstädter Molecular Modelling Workshop, Erlangen, 05/2006

Außerdem 14 Posterbeiträge (9 zum Hauptthema der Dissertation, 5 über Kooperationsprojekte) auf nationalen und internationalen Konferenzen. Bei Interesse gebe ich Ihnen gerne einen umfassenden Überblick.

Preise und Auszeichnungen

(1) MDL-CINF Scholarship Award for Scientific Excellence
CINF Section of the American Chemical Society, Chicago, USA, 03/2007

(2) Ausgewählt zur Teilnahme am Sommerkurs “Biomolecular Simulations”
European Molecular Biology Organization (EMBO), Institute Pasteur, Paris, Frankreich, 07/2006

(3) Best Student Poster Prize
Molecular Modelling 2006, Perth, Australien, 04/2006

(4) Travel Grant, MGMS German Division, 05/2006

(5) Travel Grant, MGMS Asia/Pacific Chapter, 02/2006